Computer Networks & Communications

David C. Wyld
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 10th International Conference on Computer Networks & Communications (CCNET 2023)
- 3rd International Conference on AI, Machine Learning and Applications (AIMLA 2023)
- 11th International Conference on Instrumentation and Control Systems (CICS 2023)
- 3rd International Conference on IOT, Big Data and Security (IOTBS 2023)
- 3rd International Conference on NLP & Text Mining (NLTM 2023)
- 3rd International Conference on Computing and Information Technology (COIT 2023)

Published By

AIRCC Publishing Corporation

Volume Editors

David C. Wyld
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds)
Wireilla, Australia
E-mail: dhinaharann@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-88-6
DOI: 10.5121/csit.2023.130401 - 10.5121/csit.2023.130414

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

10th International Conference on Computer Networks & Communications (CCNET 2023), February 25 ~ 26, 2023, Vancouver, Canada, 3rd International Conference on AI, Machine Learning and Applications (AIMLA 2023), 11th International Conference on Instrumentation and Control Systems (CICS 2023), 3rd International Conference on IOT, Big Data and Security (IOTBS 2023), 3rd International Conference on NLP & Text Mining (NLTM 2023 was collocated with 3rd International Conference on Computing and Information Technology (COIT 2023). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCNET 2023, AIMLA 2023, CICS 2023, IOTBS 2023, NLTM 2023, COIT 2023. Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, CCNET 2023, AIMLA 2023, CICS 2023, IOTBS 2023, NLTM 2023, COIT 2023 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCNET 2023, AIMLA 2023, CICS 2023, IOTBS 2023, NLTM 2023, COIT 2023.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

David C. Wyld, Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds) Wireilla Net Solutions, Australia

Program Committee Members

Abd Abrahim Mosslah, University of Anbar, Iraq
Abdalhossein Rezai, University of Science and Culture, Iran
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Abdelhadi Assir, Hassan 1st University, Morocco
Abdelhalim Kessal, University of Bordj Bou Arreridj, Algeria
Abderrahmane ez-zahout, Mohammed V University, Morocco
Abhay kumar Agarwal, Kamla Nehru Institute of Technology, India
Adnan Aldemir, Van Yüzüncü Yıl University, Turkey
Ahana Gangopadhyay, University of Washington, USA
Ahmad Fakharian, Islamic Azad University, Qazvin, Iran
Ahmed Farouk AbdelGawad, Zagazig University, Egypt
Ajit Singh, Patna University, India
Akhil Gupta, Lovely Professional University, India
Akhlaq Ahmad, Umm Al Qura University, Saudi Arabia
Alireza Valipour Baboli, University Technical and Vocational, Iran
Amit Mishra, Baze university, Nigeria
Anand Nayyar, Duy Tan University, Viet Nam
Andrej V. Plotnikov, University in Odesa, Ukraine
Anghelescu Petre, University of Pitesti, Romania
Anita Yadav, Harcourt Butler Technical University, India
Anouar Abtoy, Abdelmalek Essaâdi University, Morocco
António Abreu, Polytechnic Institute of Lisbon, Portugal
Antonio Muñoz, University of Malaga, Spain
Aridj Mohamed, University of Chlef, Algeria
Assem Abdel Hamied Moussa, Chief Eng Egyptair, Egypt
Athanasios V. Vasilakos, University of Agder, Norway
Attila Kertesz, University of Szeged, Hungary
B Padmaja, Institute of Aeronautical Engineering, India
B. K. Tripathy, VIT, India
Bello Abdulazeez, Ignatius ajuru university, Nigeria
Bernard Cousin, University of Rennes, France
Bibhu Dash, University of the Cumberlands, USA
Bouden Toufik, Laboratoire des ENDs (LEND), Algeria
Brahim Lejdel, University of El-Oued, Algeria
Brigitte Jaumard, Concordia University, Canada
Cagdas Hakan Aladag, University of Toronto, Canada
Chang-Wook Han, Dong-Eui University, South Korea
Chemesse Ennehar Bencheriet, University in Guelma, Algeria
Cheng Siong Chin, Newcastle University, Singapore
Chih-Hung Wang, National Chiayi University, Taiwan
Christian Mancas, Ovidius University, Romania
Christina Politi, University of Peloponnese,Greece
Claudio Cuevas, Federal University of Pernambuco, Brazil
Cristina Rottondi, Politecnico di Torino, Italy

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

10th International Conference on Computer Networks & Communications (CCNET 2023)

3rd International Conference on AI, Machine Learning and Applications (AIMLA 2023)

11th International Conference on Instrumentation and Control Systems (CICS 2023)

3rd International Conference on IOT, Big Data and Security (IOTBS 2023)

3rd International Conference on NLP & Text Mining (NLTM 2023)

3rd International Conference on Computing and Information Technology (COIT 2023)

MEASURING PERFORMANCE OF WEB PROTOCOL WITH UPDATED TRANSPORT LAYER TECHNIQUES FOR FASTER WEB BROWSING

Ziaul Hossain[1] & Gorry Fairhurst[2]

[1] Department of Computing, University of Fraser Valley, Abbotsford, BC, Canada
[2] School of Engineering, University of Aberdeen, UK

ABSTRACT

Popular Internet applications such as web browsing, web video download or variable-rate voice suffer from standard Transport Control Protocol (TCP) behaviour because their transmission rate and pattern are different from conventional bulk transfer applications. Previous works have analysed the interaction of these applications with the congestion control algorithms in TCP and proposed Congestion Window Validation (CWV) as a solution. However, this method was incomplete and has been shown to present drawbacks. This paper focuses on the 'newCWV' which was proposed to address these drawbacks. newCWV depicts a practical mechanism to estimate the available path capacity and suggests a more appropriate congestion control behaviour. These new modifications benefit variable-rate applications that are bursty in nature, with shorter transfer durations. In this paper, this algorithm was implemented in the Linux TCP/IP stack and tested by experiments, where results indicate that, with newCWV, the browsing can get 50% faster in an uncongested network.

KEYWORDS

Network Protocols, HTTP, TCP, Congestion Control, newCWV, Bursty TCP traffic

1. INTRODUCTION

With the development of the Internet, many applications have gained enormous popularity. Email, VoIP applications, File sharing etc., each have taken a share of the total Internet traffic, but the largest share is currently Web browsing applications with almost 70% of the total traffic across the Internet [1]. Web traffic uses TCP and HTTP [2] [3] protocols for request and delivery of the web page content. There had already been numerous developments across these protocols with a view to improve the performance without proposing any replacement of these standards.
Many of these updates to TCP focus on the congestion control mechanism as this technique define how much data can be transferred from the sender to receiver for an application flow. *cwnd* ensures that the sending rate of a flow is comparatively safe for the other flows that share the same bottleneck along the path between the sender and the receiver. But the focus in TCP improvements was primarily for bulk file transfers only. These modifications are not suitable for HTTP like traffic, which is 'bursty' (variable rate traffic with irregular intervals) in nature. This problem has been reported earlier and several attempts had also been made to realise a solution [4][5] . Unfortunately, these solutions were still conservative and lack proper measurement of the available path capacity to set the congestion window (*cwnd*) – the most important parameter of

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 01-16, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130401

the congestion control mechanism. This shortcoming limits the performance of bursty applications like HTTP.

A newer method has been developed termed as 'newCWV' [6]. When sending bursty or rate-limited traffic, this new method allows a sender to estimate the path capacity more accurately and set the *cwnd* to an appropriate value accordingly. The rationale for newCWV is presented briefly and the algorithm is explained elaborately in [6]. But there is a void in validating the arguments and also in measuring the expected application performance improvement with this proposal.
This paper aims to explain the motivation behind developing newCWV in detail and then analyse the web traffic transfer durations in order to measure improvements. Particular focus of this paper is on the implementation and integration of newCWV into the Linux and run experiments to support the theory. Through experiments, this paper shows that, when HTTP-like traffic uses 'newCWV', there is significant gain in performance compared to conventional TCP. Web browsing can proceed in approximately 50% faster rate in an uncongested network with the newCWV.

Section 2 of this paper explains the bursty behaviour of the HTTP traffic, the basics of TCP congestion control and the state of the art to set the background. Then, section 3 explains the modification specified in [6]. Section 4 summarises the experiment and presents the results with discussion. Finally, section 5 concludes the findings.

2. BACKGROUND

To understand the problem of transporting HTTP-like traffic with unmodified TCP, the behaviour of these protocols needs to be examined. This section explores the bursty behaviour of the HTTP protocol, the conventional congestion control of TCP and explains the interactions when these are used together.

2.1. Nature of HTTP traffic

The HTTP web traffic is naturally bursty in nature. Burstiness could be termed as a property of an application where the traffic is generated in a random manner at different rates over its running time. This could be characterised as periods of inactivity separated by periods when the chunks of data are downloaded. [7] showed that popular HTTP applications such as Web video (YouTube), Maps (Google Maps), Remote Control (LogMeIn) all send data in the downstream at variable rate with spikes up to 400KB/s, separated by periods with no activity. This burstiness is caused by the HTTP request pattern in the client/user application. Besides application behaviour, small-scale burstiness can also be caused by TCP.

[8] showed that TCP self-clocking, combined with network queuing delay (due to packets of the same flow or cross traffic) can shape the packet inter-arrivals of a TCP flow resulting in an ON-OFF pattern. With a view to modelling the inactivity (OFF) periods of the web clients, [9] showed that the OFF duration could range from a few seconds to many tens of seconds, with a probability of 80% and 10% respectively, which causes burstiness of a TCP connection when requesting content from the server.

The cited papers all agree that burstiness has become a common pattern for HTTP traffic. A simple experiment was run that captured packets while accessing a webpage from a browser to capture this bursty behaviour.

Figure 1 shows the resulting burstiness. In this capture, the chunks of data are the results of HTTP GET requests made by the client. It is visible that there are considerable inactive periods between consecutive bursts.

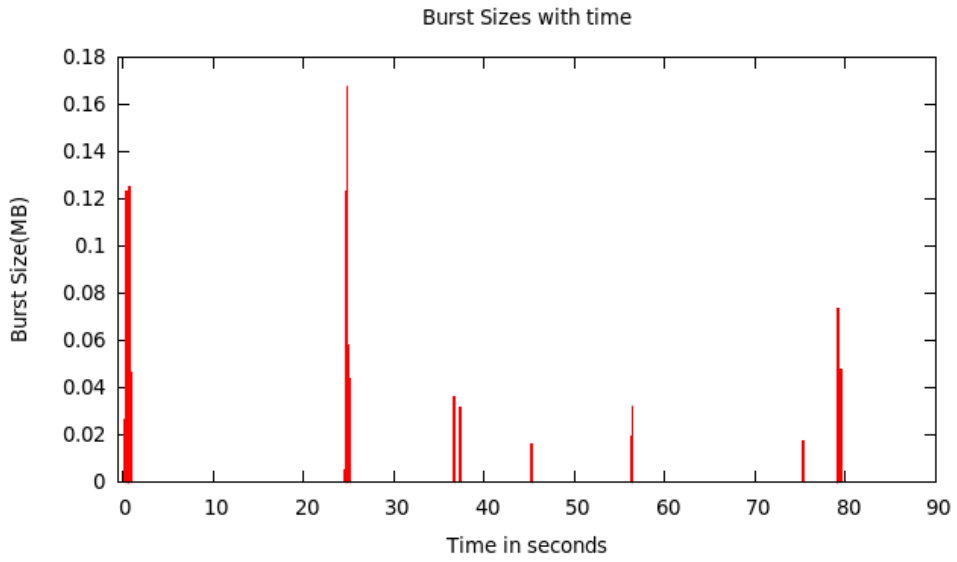


Figure 1. Bursty traffic pattern of HTTP web for a single web page

2.2. TCP Congestion Control Mechanisms

After being first standardized in 1981 by the Internet Engineering Task Force (IETF), TCP was enriched by a series of developments to face numerous challenges occurring in the underlying network. [10] provided a roadmap that described many of these changes.

A basic operating procedure of TCP is explained in the remainder of this subsection.
A TCP sender uses a parameter called the congestion window, or '*cwnd*'. This is initialised to the Initial Window (IW) size. It determines the amount of data that can be sent to the receiver while before receiving an acknowledgement from the receiver. The value of the cwnd is important, as it ultimately dictates the transfer rate and eventually the response time for an HTTP connection. TCP uses four congestion control algorithms to set the value of this ''cwnd that were specified by RFC2001, RFC2581 and RFC5681 [11][12][13]. They are Slow Start, Congestion Avoidance, Fast Retransmit and Fast Recovery.

Slow Start: In the Slow Start phase, a TCP sender sends data limited by the cwnd value and waits for Acknowledgement (ACK) packet from the receiver. Upon receiving an ACK, the value of the *cwnd* is increased by one segment. So, if a sender sent 4 segments at first (because *cwnd* = 4), and then receives 4 ACKs for these segments, then after increasing *cwnd* for each segment, the final *cwnd* value will be 6, and 6 segments can be sent. As a result of this cumulative increase, the cwnd increases using an exponential function. This continues until it reaches the Slow Start threshold (ssthresh) or the sender discovers congestion or encounters a loss.

Congestion Avoidance (CA): When the *cwnd* reaches the ssthresh, a limit is imposed on the increase of the *cwnd*. After this point, the size of the *cwnd* is only increased by one segment in one RTT. For example, if 8 segments are sent altogether, then when the 8 segments are acknowledged, the *cwnd* becomes 4 only. This corresponds to a slower linear growth of *cwnd*.

Fast Retransmit: When a packet is lost, the subsequent packets are received out of order at the receiver. When this happens, the receiver sends duplicate ACK packets when each segment is received. All the ACK packets acknowledge the same sequence number. Upon receiving the first duplicate ACK, the sender does not immediately take action, but waits to see if this is a re-ordering issue or a packet loss. When it receives a series of duplicate ACKs equal to the DupACK threshold (3 as currently standardised), the sender TCP retransmits the segment, and resets the congestion state.

Fast Recovery: When a segment is lost, rather than setting the *cwnd* to the lowest value and then send packets in sequence, it is assumed that a better approach would be to start from an intermediate value so that the flow is not badly affected. So, after a lost segment has been transmitted, CA is performed instead of Slow Start. The *cwnd* is set to (ssthresh + DupACK) segments. This is to virtually inflate the network. Since DupACKs packets have been received, this means these packets have left the network (i.e. had been received successfully). With each further DupACK, the *cwnd* is incremented by one segment. When a new ACK is received, the *cwnd* is reset to ssthresh and CA is resumed.

Selective ACK (SACK): SACK acknowledges reception of out of sequence packets. This helps avoid retransmission of already received packets. Using SACK, the receiver appends a TCP option in the DupACK header that contains a range of non-contiguous data that have been received. This allows the sender to resend only the packets that were missing from the flow. Support for SACK is negotiated at the beginning of a TCP connection; it can be used if both ends support the SACK option. [14] showed that NewReno with SACK enabled, requires fewer packet transmissions in the First Recovery phase, reduces unnecessary duplicate transmission and avoids waiting time.

2.3. TCP Variants

Different variants of TCP have evolved using combinations of these algorithms and with modifications to control the data flow and to improve response to network congestion. When a loss is detected, the TCP sender takes measures to control the flow of further packets by reducing the sending rate. Different TCP variants such as Tahoe, Reno, NewReno, which act differently in response to detected congestion.

Tahoe used Slow Start, Congestion Avoidance and Fast Retransmit. A problem with Tahoe is that restarted from the initial *cwnd* value after each packet loss. This resulted in lower throughput. To deal with this, Reno implemented Fast Recovery. This effectively recovered a of single packet loss within a window. If two or more packets were dropped in the same window, the sender was forced to timeout and restart in Slow Start. To overcome this problem, NewReno uses a modified Fast Retransmission phase based on the research [15][16]. This starts when a packet is lost and ends when a Full ACK is received, which means that all the packets transmitted between the lost packet and the last packet have been successfully received. However, if there are multiple packet-drops, then the sender will acknowledge a packet that has a lower sequence number than the last transmitted packet. This is a Partial ACK, and in this case, the lost packet is retransmitted immediately without waiting for receiving duplicate ACKs. This avoids a possible timeout. This ensures better performance than Reno, but may need to restart after a timeout if many packets are dropped from the same window.

When there are multiple losses, SACK provides better performance by enabling the receiver to inform the sender when there are multiple packet losses. A SACK block indicates a contiguous block of data that has been successfully received. The segment just before the first block and the

gap between any two consecutive blocks denote lost segments (more accurately, these are segments for which there is no acknowledgment). When the sender receives a SACK option, it can find out which segments may be lost and retransmits them. SACK is widely implemented in the current Internet, usually in combination with NewReno [10]. Therefore NewReno with SACK is considered as the standard congestion control for TCP.

Figure 2 shows the cwnd evolution for different variants with a 50ms path delay. It can be noticed in this figure that after some variation during the early stage (about 3s), all the variants reach a steady state with a similar behaviour. At time 1s, all variants start in the Slow Start phase and increase the cwnd exponentially until ssthresh is reached or after it suffers a loss. When this happens, Tahoe reduces its *cwnd* to an initial value (IW) and resumes in Slow Start. Once it reaches ssthresh, the sender enters the CA phase (linear increase of cwnd). Other variants enter Fast Retransmission and Fast Recovery phase, where the *cwnd* progresses almost linearly.

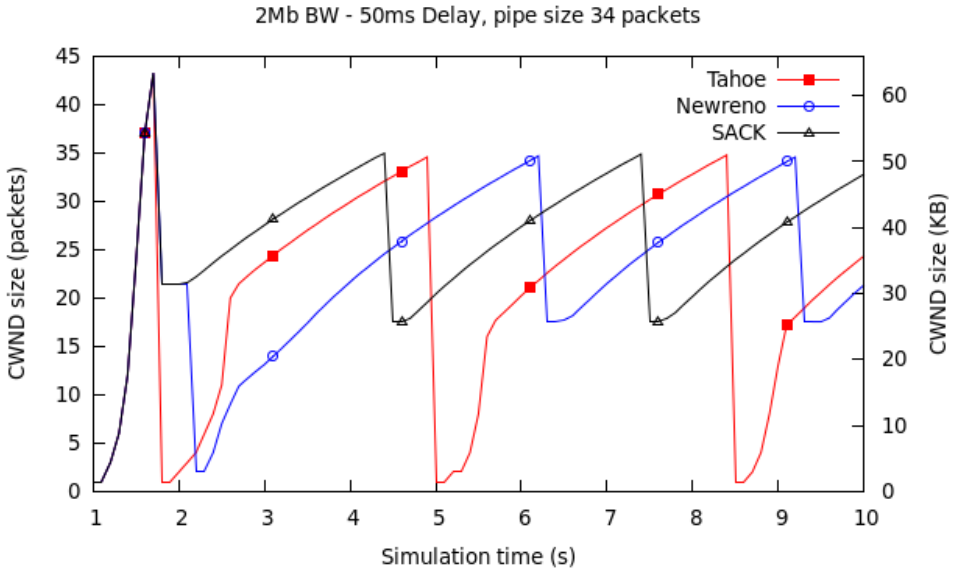


Figure 2. Congestion Window evolution with time during the simulation

Whenever packet loss occurs (at 1.8s, 5s and 8.5s for Tahoe) Tahoe enters the Fast Retransmit phase and retransmits the lost packet. The *cwnd* is set to the restart window (RW) (normally 1 segment) and continues in the Slow Start phase.

NewReno and SACK enter the Fast Recovery stage (at 1.8s for the first loss) where the ssthresh is set to a new value that is half the size of the unacknowledged data and then *cwnd* is set to ssthresh. At 2s, to more losses cause NewReno to fail to recover and eventually the sender is forced to enter the Slow Start phase. SACK was successful in recovering and eventually followed the CA until it faced another loss. Later, during the simulation, NewReno successfully recovered using Fast Retransmit and Fast Recovery. Then it moves to the CA phase.

2.4. TCP CWV

Standard TCP congestion control required that when an application is idle for a period greater than the Retransmission Timeout (RTO), the *cwnd* is reset to a small value. So, the next burst of data requires the sender to re-enter the Slow Start phase from this small value. Several RTTs may

be consumed before the previous sending rate is again achieved. This approach is too conservative, in that it fails to use available capacity. For a bursty application, this scenario is quite common where each burst is separated by an idle period. As a result, the application performance suffers from this conservative behaviour of TCP.

Figure 3 explains the situation as a diagram. After RTO, the cwnd (red bold solid line) drops to the RW. Then it takes long time to grow back for the next burst. So, the net burst is unnecessarily delayed while the path capacity might have been enough to transmit the burst in shorter RTTs.

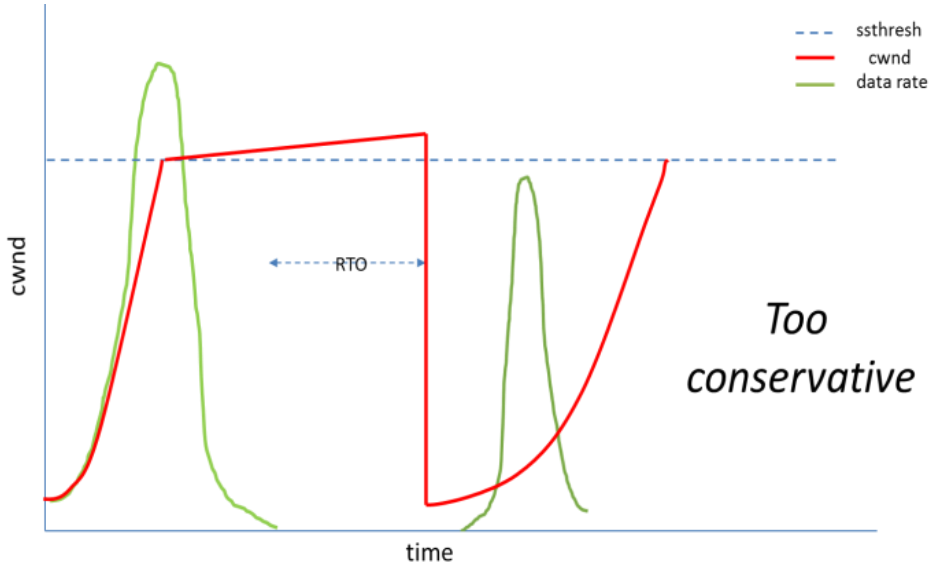


Figure 3. Reducing the cwnd to a low value of RW makes it overly conservative for idle period

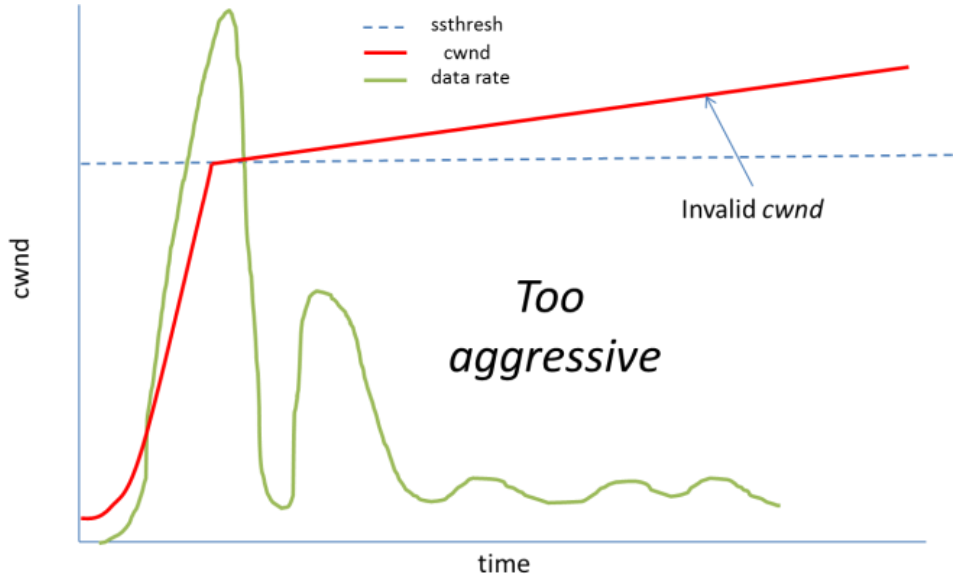


Figure 4. Increasing the cwnd during the application limited period makes it invalid

On the other hand, during an application-limited period, a Standard TCP sender continues to grow the *cwnd* for every received acknowledged packet (ACK), allowing the cwnd to reach an

arbitrarily large value. However, in this case the packet probes along the transmission path are sent at a lower rate than permitted by *cwnd*, so the reception of an ACK does not actually provide evidence that the network path was able to sustain the transmission rate reflected by the current cwnd. The *cwnd* is called 'invalid'.

Figure 4 explains such a scenario. The actual path capacity which may be significantly lower than the *cwnd*, can be mistaken.

If an application with an invalid cwnd were to suddenly increase its transmission rate, the sender would be allowed to immediately inject a significant volume of additional traffic into the network. This could lead to severe network congestion, potentially harming other flows that share a common bottleneck.

TCP Congestion Window Validation (TCP-CWV), was first specified in RFC 2861 [4], was proposed as an experimental standard by the IETF. The intention was to find a remedy for the problems imposed by TCP when used by a bursty application. TCP-CWV changed how cwnd is updated and is to be used during an idle or application-limited period.

TCP-CWV modified the congestion control algorithm of standard TCP during an application-limited period when the cwnd had not been fully utilised for a period larger than an RTO.
During an idle period, which is greater than one RTO, TCP-CWV reduced cwnd by half for every RTO period. This is equivalent to exponentially decaying cwnd during the idle period compared to reducing the cwnd in a single step with standard TCP. This is common traffic pattern for bursty applications to have an idle period in the order of seconds – which could be larger than a few RTOs worth of time. As a result, TCP-CWV ultimately reduces to RW and causes problem like standard TCP.

Another recommendation of CWV was to set the *cwnd* according to (*w_used*+*cwnd*)/2 for each RTO period that does not utilise the full *cwnd*, where *w_used* is the maximum amount that has been used since the sender was last network-limited/cwnd-limited. This avoids a growth of cwnd to an invalid value; it can cause the cwnd to reduce to a value that is close to the current application rate.

This results in two problems:
First, the *cwnd* should reflect the network capacity for a flow and control the amount of data that the network could sustain. However, CWV tends to set the cwnd according to the traffic pattern and application rate - only seeking to be conservative in use of network capacity. As a result, the *cwnd* is set to a lower value that is more conservative than when using standard TCP, which would have allowed larger bursts.

Secondly, CWV used w_used, the amount of data that has been sent by the application, but not yet acknowledged. In an application-limited period where the application is not using the allowed path capacity, w_used is does not reveal the available capacity. According to this approach, the *cwnd* is set to a value that is determined by the application's sending rate in the last RTO period (last few RTTs), rather than the network capacity. This impacts the application performance where the subsequent bursts are to be rate-limited and would take longer to complete. So, this should not be regarded as the available path capacity for the TCP flow that is recoded in the *cwnd*.

In summary, when TCP-CWV was specified in 2000, it identified a need to change the way TCP responded for bursty applications but failed to offer a complete solution.

3. TCP NEWCWV: MODIFICATION FOR HTTP-LIKE TRAFFIC

When newCWV was standardised in 2015, it introduced a variable called 'pipeACK' that was used to measure the acknowledged size of the network pipe. The pipeACK variable is considered as a safe bound for the capacity available to the sender since this represents the actual amount of data that was successfully transmitted in an RTT from the sender to the receiver. This variable can be computed by measuring the volume of data that have been acknowledged by the receiver within the last RTT.

The pipeACK is used to determine if the sender has validated the cwnd. The sender enters the non-validated phase when:

$$pipeACK < \frac{1}{2} \times cwnd$$

newCWV also defined a new phase. A sender was allowed to use the cwnd for a period (5 minutes), called the Non-Validated Period (NVP). During the NVP, the *cwnd* is preserved. The reason for storing the *cwnd* for several minutes because it is the default server timeout for TCP connection.

In summary, newCWV brought stability for both phases of rate-limiting period and application limiting period for HTTP like traffic. An algorithm was proposed and implemented in the Linux Kernel module, which was used to verify the effectiveness of this modification in the next section.

4. EXPERIMENT, RESULTS & DISCUSSION

This section first describes the network emulation used to explore the behaviour of newCWV. Then presents the findings in different scenarios with possible explanations for such behaviour.

4.1. Experimental Setup

A network emulation method was chosen to conduct the experiments because this enables a real implementation of network protocols to be tested in a controlled environment. The test bed used a dumbbell topology representing a single network path bottleneck (refer to Figure 5).

Client 1 and Server 1 were used to benchmark the newCWV behaviour for the main traffic (either HTTP web or HTTP streaming content). Another server (client 2, server 2) was used to inject cross traffic (in this case a large file transfer using FTP) into a shared network bottleneck. All servers ran Linux kernel versions 3.12, and the clients were running 3.8 or greater.

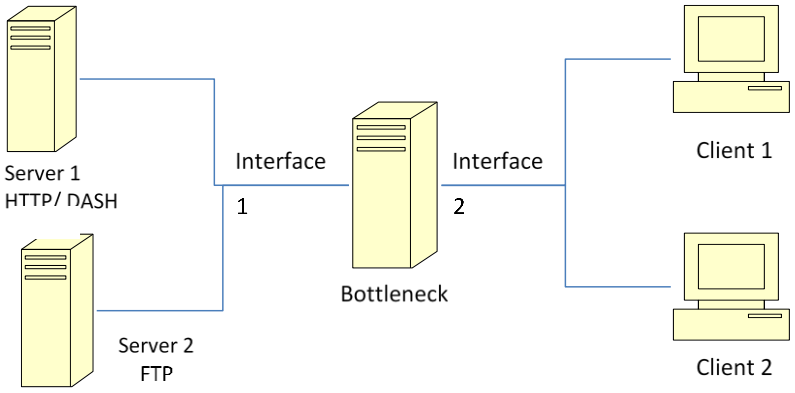


Figure 5. Experiment topology: the bottleneck router imposed a fixed bandwidth and delay between the client and server

Server 1 acts as a web server or streaming server that used standard TCP (NewReno with SACK). It was installed with the newCWV Loadable Kernel Modules (LKM) for Linux, the traffic generators and iproute2 utilities (to enabling pacing when required) allowing this to be chosen at the start of each experiment.

The experiments are run across a range of time intervals that represent values that range between HTTP response bursts (idle periods) and for HTTP response sizes larger than a particular value (Burst size after idle). The results obtained from multiple iterations of these experiments are averaged to measure the completion time of the HTTP/TCP connections for different combinations of idle periods and burst sizes.

The comparison plots, shown in the results section, present the ***improvement in burst transfer time*** (less time required for transmission) when newCWV is used compared to using a standard TCP (NewReno with SACK). The performance gain in transfer time (% improvement) is calculated by taking an average of the transfer gain over all bursts. The transfer gain was calculated by the following formula, where time taken in NewReno/SACK is ***Tr*** and time taken for a burst with newCWV is ***Tc***:

$$\text{Gain (in percentage)} = (Tr - Tc) / Tr \times 100$$

The gain can be positive where the burst is transmitted faster or negative when a particular HTTP response takes longer to transmit due to loss. A positive average of all these values indicates an overall gain in performance – the higher the value, the better.

The table below (Table 1) summarizes the experiment parameters:

Table 1: Experiment Parameters

Parameter	Value
TCP Initial Window (IW)	3 Segments
Ssthresh sharing	NO
Bottleneck Bandwidth	2 Mbps
Delay / RTT	200 ms
HTTP Generator	Tmix tool
Linux Kernel	3.12
No of HTTP connections	3151
Total Data analysed	7.68 GB
Average Transfer rate	700 kbps
Iterations with same parameters	5

4.2. Comparing performance over an uncongested path

To understand the effect of newCWV in a non-congested scenario, experiments were run with no bandwidth limit at the bottleneck router; only a link delay of 200 ms was applied. There was no cross-traffic and no rate limit was applied. Figure 6 below presents the performance improvement of newCWV compared to NewReno, plotted burst sizes vs. different idle periods.

The improvement is visible in this figure (Figure 6). A newCWV sender transfers a burst in 37-62% less time than NewReno. Larger improvements are achieved for the higher burst sizes, as expected; about 10% more improvement is achieved for bursts of 80KB (60%) than 5KB bursts (50%). While standard TCP reduces its cwnd after an idle period, newCWV retains a larger cwnd and is able to transfer the burst in less time, saving several RTTs – an approximate average of 50% improvement suggests newCWV requires half the RTTs compared to NewReno.

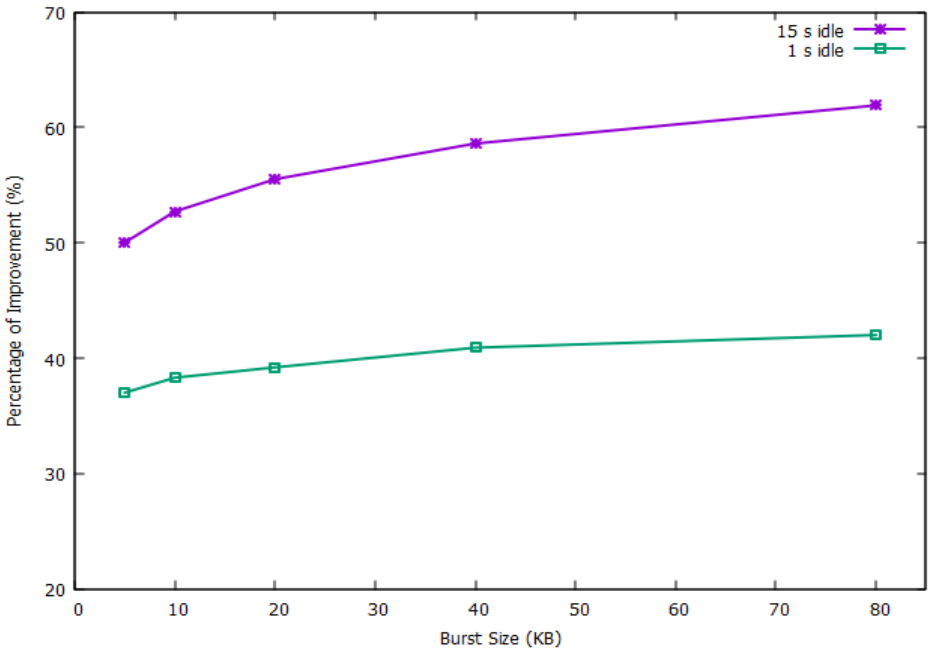


Figure 6. Performance Improvement of HTTP traffic shown when newCWV is used instead of NewReno over different burst sizes and idle periods.

Another interesting fact is that for a same size of burst, for example a 40 KB burst, would encounter almost 20% more improvement in performance when the idle period is larger. So, it shows that for real-life web browsing traffic, even if the idle period is high newCWV will support more traffic than the conventional TCP.

In short, for an uncongested scenario (as may be expected in a LAN), newCWV shows improved performance over standard TCP.

4.3. Comparing performance in a congested path

To test the effectiveness of newCWV in an Internet context, a bottleneck of 2 Mbps was set with a finite router buffer (30 KB). The path MTU was 1500 B, which ensured a maximum of 20 segments to be queued in the buffer. The newCWV protocol still shows improvement over standard TCP, which now varies from 10-35% over the idle period – burst size domain (shown in Figure 7).

While in the previous scenario, there were no other traffic, the performance improved more for higher burst sizes. However, in this congested scenario, the trend is somewhat opposite: Higher burst sizes offer less improvement. In the case of the idle period comparison, the similarity remains, where a larger idle period increases the improvement as in the previous non-congested scenario.

For bursts larger than 5 KB (larger than an IW of 4 KB), it takes about 25-33% less time on average for a transfer with an idle period. A larger improvement is demonstrated around 35% with newCWV, but the advantage diminishes for larger burst sizes (for 40 KB or 80 KB), because it encounters higher loss. The newCWV sender is prone to a higher loss rate for larger bursts. These bursts can appear either at the beginning of the TCP connection or after an idle period.

Figure 8 confirms that the number of losses is higher when using newCWV.

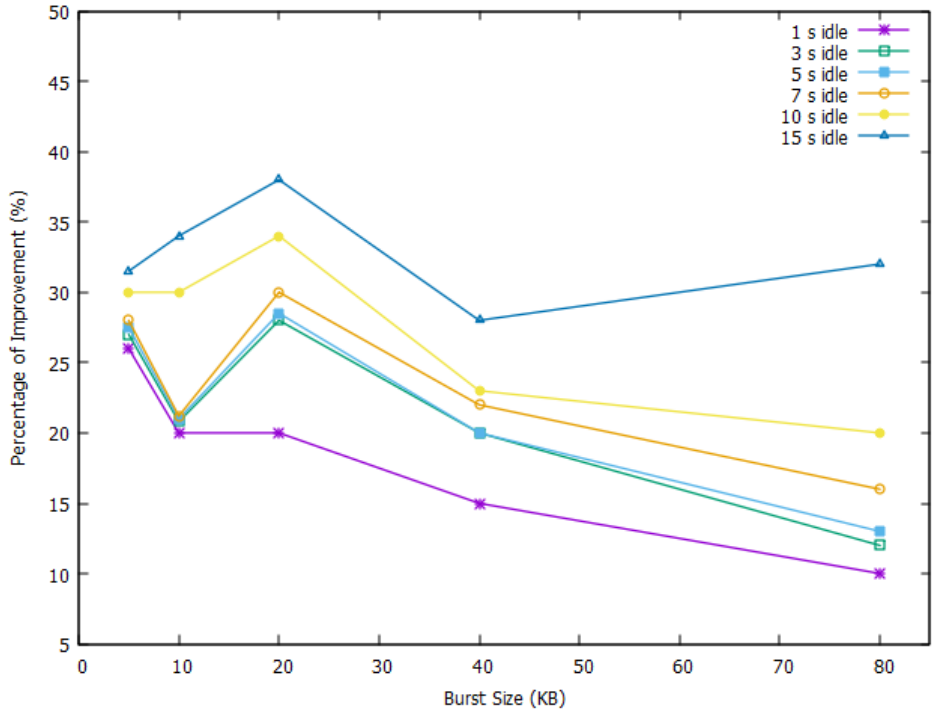


Figure 7. HTTP performance improvement in percentage is shown in a congested scenario when newCWV is used instead of NewReno over different burst sizes and idle periods.

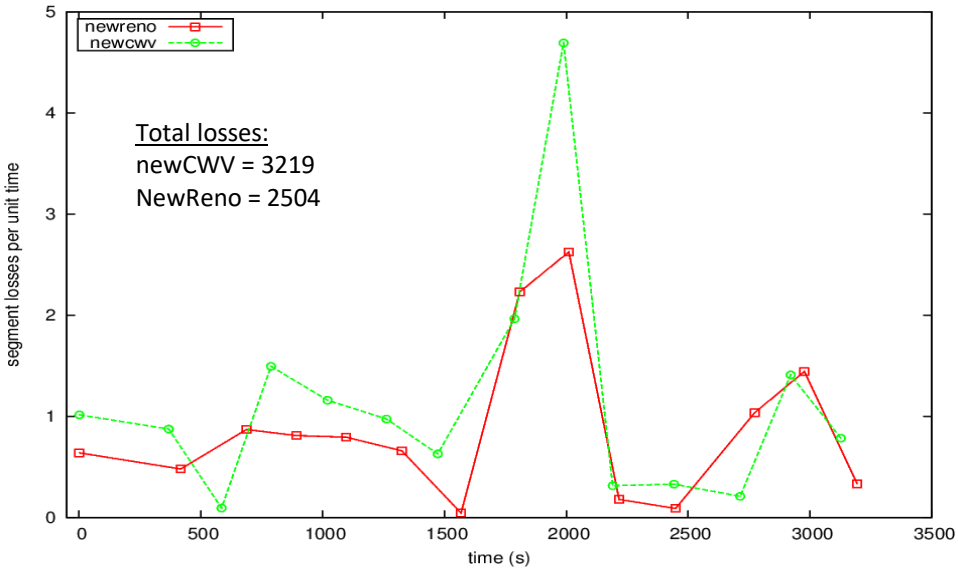


Figure 8. Loss plot comparing NewReno and newCWV; newCWV suffers more loss on average than standard TCP.

A newCWV sender allows larger bursts into the network for a large HTTP response. With a finite network buffer this will increase the probability of (burst) loss and queuing delay for this flow and other flows that share a common bottleneck (e.g., higher packet loss and jitter for concurrent real-time applications).

Although newCWV continues to show better performance in delivering bursts faster in a congested scenario, higher burst losses for larger bursts may degrade the overall average improvement in burst transfer times that could have been possible for HTTP flows.

4.4. Effect of other application on HTTP with newCWV

At first, the HTTP workload was run with newCWV without pacing. Table 2 shows that, for HTTP responses with a size of 5KB or more, there was an improvement in transfer times of 18-20%, although it is low compared to the previous cases without any cross traffic.

For large responses, such as 80 KB or more, the performance of newCWV reduces compared to standard TCP. The newCWV sender was observed to take about 10-15% more time to complete the bursts than NewReno. The large level of packet loss (and therefore delay) caused by large bursts being injected into the network eliminated the benefit of newCWV. This indicates that some burst mitigation technique is desirable.

Table 2. Performance Improvement in Percentages when HTTP runs with newCWV against NewReno. A negative value means performance degradation.

Burst Size (KB)	**Idle Periods**					
	1 s	**3 s**	**5 s**	**7 s**	**10 s**	**15 s**
5	17.9 %	18.3 %	18.4 %	18.9 %	19.1 %	19.5 %
10	10.2 %	10.6 %	10.7 %	11.2 %	11.4 %	11.7 %
20	6.7 %	6.7 %	6.9 %	7.1 %	7.3 %	7.3 %
40	4.2 %	4.4 %	4.4 %	4.6 %	4.9 %	5.2 %
80	-10.2 %	- 12.5 %	- 13.4%	- 13.9 %	- 15.1 %	- 15.3 %

In summary, when using a very congested bottleneck shared with other applications, newCWV needs to be combined with pacing – sending the burst in regular intervals – to ensure a performance improvement. Otherwise, it can lead to significant loss and induce delay to the applications using the bottleneck.

To assess the effect of newCWV, an FTP application (running NewReno) shared the bottleneck with a HTTP workload using different algorithms: NewReno, newCWV and paced newCWV.

Figure 9, shows that for the whole period of the experiment (about an hour), the FTP application competed with the HTTP traffic for a share of the capacity of the 2 Mbps bottleneck. Fluctuations in FTP throughput were observed as it shared the bottleneck with the variable rate HTTP web traffic. FTP did not suffer from starvation when the other TCP was using newCWV.
The curves for newCWV follow the curve for NewReno with hardly any differences. Though newCWV seems to be more aggressive after an idle period than standard TCP (NewReno), which helps a bursty sender application, it was reacting to congestion appropriately sharing the bottleneck with another long-lived TCP flow. This depicts the friendliness of newCWV with other TCP application like FTP.

In summary, newCWV demonstrated improved performance for HTTP traffic in both a congested and uncongested scenario. It is recommended that newCWV is used in combination with pacing, to smooth out the burst and hence also to reduce losses. newCWV is also fair in a sense that it does not poses significant threat (aggressiveness or starvation) to other co-existing TCP flows.

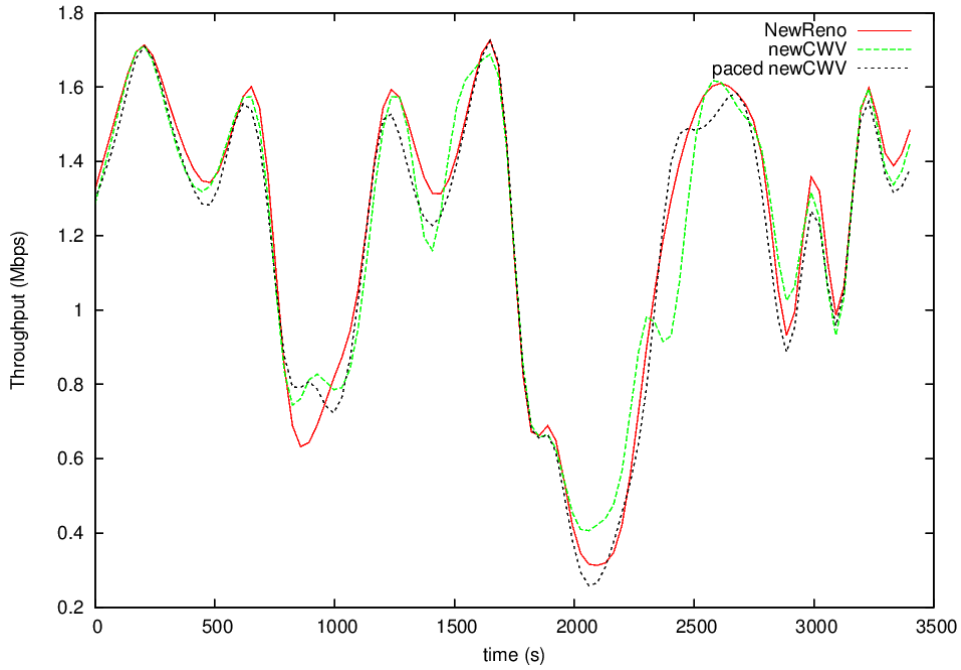


Figure 9. FTP cross-traffic throughput; no significant differences in background application performance while the HTTP traffic was running different algorithms.

4.5. Discussion

Since newCWV can avoid suboptimal performance, by defining a new way to use the *cwnd* and *ssthresh* during a rate-limited interval and specifies how to update these parameters after congestion has been detected. The mechanism defined in RFC 7661 is considered safe to use even when *cwnd* is greater than the receive window [17], because it validates the *cwnd* based on the amount of data acknowledged by the network in an RTT, which implicitly accounts for the allowed receive window.

The paper evaluated a working version of this algorithm in Linux. Since newCWV was published as an experimental specification in the RFC-series as RFC 7661, it has been implemented in some production endpoint TCP stacks. It is referenced in the latest IETF QUIC [18] transport specification: QUIC Loss Detection and Congestion Control, (RFC 9002). It is also referenced in a range of other IETF specifications, that includes Self-Clocked Rate Adaptation for Multimedia (RFC 8298), Model-Based Metrics for Bulk Transport Capacity (RFC 8337), TCP Control Block Interdependence (RFC 9040) and Operational Considerations for Streaming Media (RFC 9317).

5. CONCLUSION

Web-based traffic is the dominant type of traffic in today's Internet. As web uses HTTP/2, that uses TCP as underlying protocol, it is very important to study the transport behaviour to ensure the browsing can be made faster. A set of problems have been identified by earlier research works when bursty HTTP application use traditional TCP congestion control. Although some solutions had been proposed, they were limited and did not properly address the key requirements. newCWV seeks to address the congestion control problems and is implementable.
This paper found that the newCWV mechanism is useful for applications with variable rates in both rate-limited periods and idle periods. newCWV can lead HTTP based traffic to completion in a 50% faster manner, which means web browsing will be much more faster, web based video

streaming would be some more smoother etc. Moreover, it does not induce any harm to other network traffic sharing a common bottleneck.

The great impact of using newCWV is that application designers do not have to worry about the underling transport support for bursty applications, since the transport can accommodate a wide range of traffic variation. This gives application developers more freedom when developing new applications and can encourage the development of next generation Internet applications. For future work, it would be interesting to see the performance comparison with current TCP and QUIC implementations and to consider a variety of other network conditions.

ACKNOWLEDGEMENTS

The author acknowledges the Electronics Research Group of University of Aberdeen, UK, for all the support in conducting these experiments. This research was completed as a part of the University of Aberdeen, dot.rural project. (EP/G066051/1).

ACRONYMS

ACK	Acknowledgement
cwnd	congestion window
CA	Congestion Avoidance
CWV	Congestion Window Validation
DASH	Dynamic Adaptive Streaming over HTTP
FTP	File Transfer Protocol
HTTP	Hyper-Text Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
IW	Initial Window
LKM	Loadable Kernel Module
MTU	Maximum Transfer Unit
NVP	Non-Validated Period
TCP	Transmission Control Protocol
RFC	Request for Comments
RTO	Retransmission Time Out
RTT	Round Trip Time
RW	Restart Window
SACK	Selective ACKnowledgement
ssthresh	Slow Start Threshold

REFERENCES

[1] Sandvine Resources, (2022), “Global Internet Phenomena Report”, Whitepaper, Sandvine Corporations.
[2] Fielding R. *et al* (1999), “Hypertext Transfer Protocol -- HTTP/1.1”, RFC 2616, IETF
[3] ISI (1981), “Transmission Control Protocol”, RFC 793, IETF.
[4] Handley, M., Padhye, J. and Floyd, S. (2000), “TCP Congestion window Validation”, RFC 2861, IETF.
[5] Biswas, I. (2011), “Internet congestion control for variable rate TCP traffic”, PhD Thesis, School of Engineering, University of Aberdeen.
[6] Fairhusrt, G., Sathiaseelan, A. & Secchi, R. (2015), “Updating TCP to Support Rate-Limited Traffic”, RFC 7661, IETF.
[7] Augustin, B. and Mellouk, A. (2011), “On Traffic Patterns of HTTP Applications”, Proceedings of IEEE Globecomm, Houston, USA.

[8] Jiang, H. and Dovrolis, C. (2205), “Why is the internet traffic bursty in short time scales?”, Proceedings of the ACM SIGMETRICS international conference on Measurement and Modeling of Computer Systems, Banff, Canada.
[9] Casilari, E. et. al. (2004), “Modelling of Individual and Aggregate Web Traffic”, Proceedings of 7th IEEE conference in High Speed Networks and Multimedia Communications, Toulouse, France.
[10] Duke, M., et al (2015) A Roadmap for Transmission Control Protocol (TCP) Specification Documents, RFC 7414, IETF
[11] Stevens, W. (1997), "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC 2001, IETF.
[12] Allman, M., Paxson, V. and Stevens, W. (1999), "TCP Congestion Control", RFC 2581, IETF.
[13] Allman, M., Paxson, V. and Blanton E. (2009), “TCP Congestion Control”, RFC 5681, IETF.
[14] Fall, K. and Floyd, S. (1996), "Simulation-based Comparisons of Tahoe, Reno and SACK TCP", ACM SIGCOMM Computer Communication Review, vol. 26, no. 3, pp. 5-21.
[15] Hoe, J. (1996), “Improving the Start-up Behavior of a Congestion Control Scheme for TCP”, Proceedings of the ACM SIGCOMM.
[16] Floyd, S., Henderson, T. and Gurtov, A. (2004), “The NewReno Modification to TCP's Fast Recovery Algorithm”, RFC 3782, IETF.
[17] Xu, L., et al, CUBIC for Fast and Long-Distance Networks, draft-ietf-tcpm-rfc8312bis, Work-In-Progress, TCPM Working Group, IETF.
[18] Iyenger, J. & Thomson, M. “QUIC: A UDP-Based Multiplexed and Secure Transport”, RFC 9000, IETF.

AUTHORS

Dr Ziaul Hossain is a Network Researcher, and his interest lies in making the web faster. He has achieved BSc in Computer Science from BUET (Bangladesh) and then pursued his MSc degree at the Queen Mary, University of London (UK). He received PhD in Engineering from University of Aberdeen (UK) where his research focused on performance of different applications over the satellite platform. He has taught at several universities and currently working as a Lecturer at the University of Fraser Valley, British Columbia, Canada.

Dr Gorry Fairhurst is a Professor in the School of Engineering at the University of Aberdeen, Scotland, UK, His research is in Internet Engineering, and broadband service provision via satellite. He is an active member of the Internet Engineering Task Force (IETF) where he chairs the Transport Area working group (TSVWG) and contributes to the IETF Internet area. He has authored more than 25 RFCs with over 150 citations within the RFC-series, and currently is contributing to groups including: QUIC, 6MAN, TSVWG, and MAPRG.

LEON: LIGHT WEIGHT EDGE DETECTION NETWORK

Nasrin Akbari and Amirali Baniasadi

Department of Computer Engineering, University of Victoria, Victoria, Canada

ABSTRACT

Deep Convolutional Neural Networks (CNNs) have achieved human-level performance in edge detection. However, there have not been enough studies on how to efficiently utilize the parameters of the neural network in edge detection applications. Therefore, the associated memory and energy costs remain high. In this paper, inspired by Depthwise Separable Convolutions and deformable convolutional networks (Deformable-ConvNet), we aim to address current inefficiencies in edge detection applications. To this end, we propose a new architecture, which we refer to as Lightweight Edge Detection Network (LEON). The proposed approach is designed to integrate the advantages of the deformable unit and DepthWise Separable convolutions architecture to create a lightweight backbone employed for efficient feature extraction. As we show, we achieve state-of-the-art accuracy while significantly reducing the complexity by carefully choosing proper components for edge detection purposes. Our results on BSDS500 and NYUDv2 demonstrate that LEON outperforms the current lightweight edge detectors while requiring only 500k parameters. It is worth mentioning that we train the network from scratch without using pre- trained weights.

KEYWORDS

Edge detection, lightweight neural network, Receptive field, network pruning

1. INTRODUCTION

Edge detection is the process of finding meaningful transitions in an image. This is done by detecting discontinuities in texture, colour, brightness, etc. Edges provide boundaries between different regions in the image. Detecting these boundaries is the first step in many computer vision tasks, such as edge-based face recognition, edge-based target recognition, scene understanding, image segmentation, fingerprint matching, license plate detection, object proposal, and object detection [1].

Edge detection is widely used in a variety of applications, including fingerprint recognition in mobile devices, well-localized maps of satellite images to suppress noise and produce realistic edge maps [2], self-driving vehicles to set the steering wheel angle based on the picture of the road [3], and finding pathological objects in medical images [4]. So, it's important to pay close attention to making a neural network that works well for the implementation.

The emergence of deep learning techniques has greatly promoted edge detection research over the past few years. Traditional approaches to the BSDS500 dataset often achieve a 0.59 ODS F-measure. DL-based methods, on the other hand, can achieve a 0.828 ODS [5]. Although recently proposed architectures achieve high accuracy, they are computationally inefficient. This makes developing lightweight networks that reduce the number of parameters while maintaining the detection accuracy critical. Figure 1 shows both the detection accuracy and complexity (model

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 17-27, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130402

size) of several well-known deep learning-based methods. As shown in figure 1, the orange dot indicates how well our model matches human perception in terms of accuracy with a few parameters.

Many deep learning-based edge detectors use VGGNet (Visual Geometry Group) [6] as their feature-based extractor because of its excellent performance. However, VGGNet has a pretty extensive backbone and employs a large number of parameters, which makes it appropriate to fit more complex tasks such as image segmentation and object recognition. This work is motivated by the fact that edge detection is a low-level image-processing task and does not require complex networks for feature extraction.

To decrease the number of parameters and floating point operations (FLOPs), we take advantage of depthwise separable convolutions [7] which disentangle the spatial and channel interaction that is mixed in a regular convolution operation. However, it reduces the performance in comparison to conventional convolution. To compensate for the reduced performance, we increase the receptive field by carefully choosing proper lightweight components for edge de- tection purposes. We explain the details in section 3.

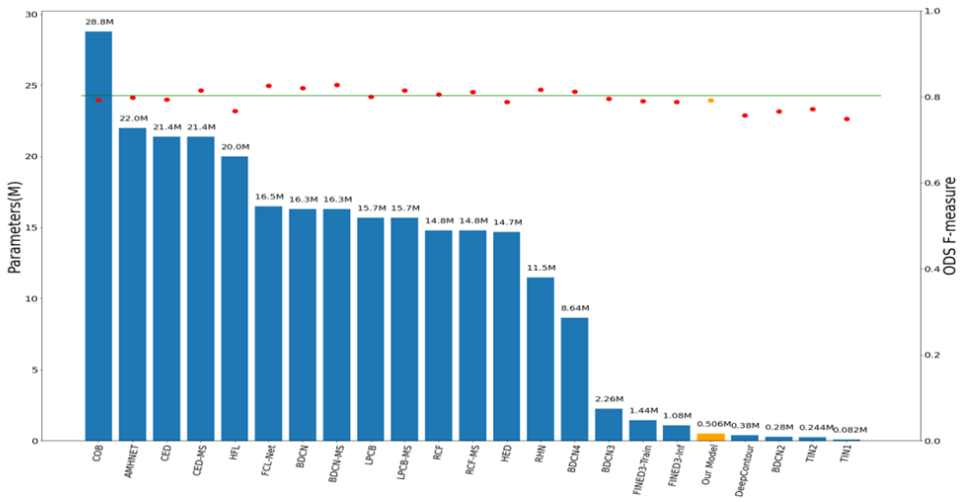


Figure 1. Comparison of complexity and accuracy performance among various edge detection schemes. Our proposed methods (orange).

The rest of this paper is organized as follows. Section 2 reviews related works and their issues. The proposed network architecture is described in section 3. Section 4 presents experimental results and compares them to the state-of-the-art edge detector networks using (Berkeley Segmentation Dataset 500) BSDS500 [8] and NYUDv2 [9] datasets. In section 5, we offer concluding remarks and discuss future research directions.

2. RELATED WORK

Over the past few years, a number of edge-detection solutions have been developed. Almost all edge detection approaches can be generally categorized into three groups, traditional edge detection, learning-based ones using handcrafted features, and deep learning networks. In the following paragraphs, we review some techniques that have been developed in recent years.

Intensity and colour gradients were the main focus of early pioneering edge detection methods. The Sobel [10] operator measures the 2-D spatial gradient of an image, emphasizing regions of

high spatial frequency that correspond to edges. The Canny algorithm [11] is a multistage edge detector. In this algorithm, the intensity of the gradients is computed by employing a filter based on the derivative of a Gaussian. The Gaussian filter reduces the impact of image noise. Subsequently, by removing non-maximum pixels of the gradient magnitude, possible edges are decreased to 1-pixel curves. Finally, applying the hysteresis threshold to the gradient magnitude, edge pixels are kept or eliminated. Zero-crossing theory based algorithms are proposed by [12, 13]. Traditional approaches suffer from some limitations, including merely focusing on the changes of local intensity while failing to recognize and remove the non-edge texture.

The introduction of learning-based edge detectors made it possible to partially overcome challenges such as texture detection problems in traditional approaches. In this group of detectors, hand-craft features are initially extracted. Later, classifiers trained using these features are applied to identify edges. The first data-driven approaches were proposed by Konishi et al. [14] who used images to learn the probability distributions of responses that correspond to the two sets of edge filters. In another work [15], random decision forests were applied to show the structure presented in local image patches. The structured forest uses colour and gradient features to high-quality output edges.

The aforementioned techniques are developed according to handcrafted features, which mostly fail to provide high-level information for semantically meaningful edge detection and have a limited capability of capturing edges at different scales. To address these issues, a number of CNN-based algorithms with strong learning capabilities have been proposed in recent years. One of the most influential in DNN-based edge detection is HED[16]. This study uses fully convolutional neural networks and deeply supervised nets to find the edge probability for every pixel. HED uses VGGNet [6] for the feature extraction and fuses all the side outputs of VGGNet features to minimize the weighted cross-entropy loss function. Since then, various extensions based on HED and VGGNet have been developed, including CED [17], AMHNet [18], RCF [19], LPCB [20], and BDCN [21].

While CNN is a very successful model, it often requires high computational power and resources. Hence, the current trend is to design efficient CNN structures that overcome such issues. Fined [22], dense extreme inception network [23], and TIN[5] have proposed a lightweight architecture for edge detection. Although these networks are light and fast, they have low detection accuracy. To achieve a better trade-off between accuracy and efficiency for edge detection, we need to optimize the architecture and initial parameters of deep learning models so that they consume fewer resources while maintaining accuracy. In this paper, we build our model by simplifying the backbone for feature extraction and carefully choosing the proper components. Therefore, we achieve good edge quality with a much simpler model compared to other studies.

3. LIGHTWEIGHT EDGE DETECTION NETWORK

Inefficiency of the models outlined in the previous section at once. In Figure 2 we present LEON's architecture. We trained the network from scratch. Below, we review the components used by LEON.

3.1. Efficient Backbone

Most deep learning-based edge detectors [17–21] employ VGGNet as their feature extraction backbone. However, we believe that edge detection is a simple task and does not need to have an extensive backbone. We reduce the backbone's complexity while keeping its efficiency by using lightweight components. To resemble the pyramid structure, we stack up three stages and use a

max-pooling operation for down sampling the features between the stages. The dimension of the output feature maps decreases as we proceed. As we move forward in the stages, the patterns get more complex; hence, there are larger combinations of patterns to be captured. Therefore, we increase the feature channel number (the number of filters) in subsequent stages to capture as many combinations as possible. Stages 1, 2, and 3 have channel numbers 16, 64, and 256, respectively. The backbone is made of mainly a combination of deformable and customized depthwise separable convolutions. To create the fused output, we use standard bilinear interpolation to up sample the low-resolution features. Then, we concatenate all the stage outputs together to form the fused output. We next elaborate on the layers and components used by LEON in detail.

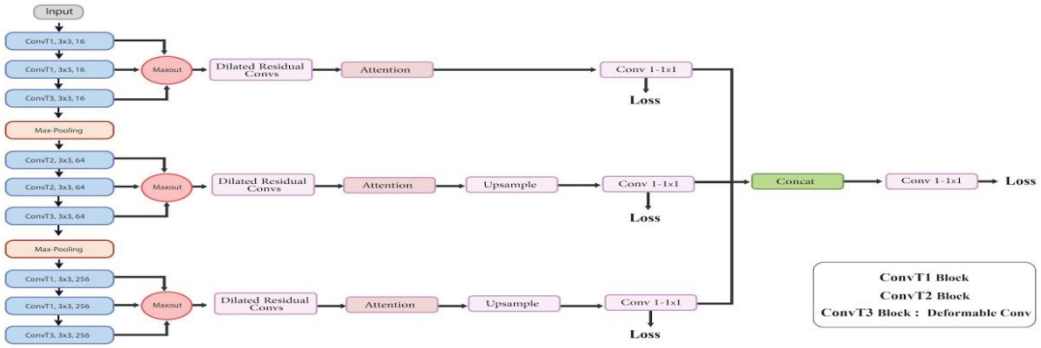


Figure 2. LEON architecture

Figure 3. Convt1 block - Convt2 block

3.1.1. Deformable convolution

Geometric transformation and variations widely existing in natural images make feature extraction tasks challenging. Standard convolution kernels have a fixed structure and have limitations in capturing geometric transformations. Deformable convolutions can address this issue efficiently. This type of convolution has the ability to change its kernel shape and the parameters within it to adapt to the image content. This adds 2D offset kernels to the regular sampling location in the standard convolution, which enables the network to have different receptive fields according to the scale of the objects. These 2D offset kernels are learnable from the preceding feature maps using additional convolutional layers and can be trained end-to-end using normal back propagation functions. We simply add this module at the end of each stage to keep our network light in terms of parameters and computation. We can strengthen our features this way before transferring them to the next stage [24].

3.1.2. Depthwise Separable Convolution

Conventional convolution performs the channels and spatial-wise computation in one step, while Depthwise Separable Convolution reduces the number of parameters by splitting the computation into two steps: 1) depthwise convolution, which applies a single convolutional filter per input channel, and 2) pointwise convolution, which creates a linear combination of the output of the depthwise convolution [7]. This approach, however, degrades accuracy. To address this problem, we reinforce the features by using additional side blocks while keeping the number of parameters as low as possible. We use RELU activation after each pointwise convolution to add non-linearity to the model for making complex decisions (Figure 3 - Convt1). To increase the accuracy of the model while keeping the number of parameters low, we modified Convt1 to Convt2 by adding pointwise convolution, which uses only a 1×1 kernel to iterate through every single point between two RELU activations. In addition, to overcome the overfitting problem, after each RELU activation, we employ a batch normalization technique as a regularizer.

3.2. Efficient Side Structure

3.2.1. Maxout Layer

At each stage, before transferring the inputs to the side output layers (from left to right), we do a Maxout operation instead of the standard concatenation block. Maxout activation can reduce the number of parameters significantly in comparison to the classical dense blocks. Instead of stacking the output of previous layers at each stage on top of each other, we only keep the maximum value at each position by inducing competition between feature maps and accelerating network convergence.

3.2.2. Dilated Residual Convolution Module

To enhance the extracted features by depth-wise separable convolution in the backbone, we connect every feature extraction layer to the dilated convolution module adopted in [5]. We use different dilation sizes to capture different levels of receptive fields in the image. The first dilation is 4, followed by 8, 12, and 16, and all the layers have 32 filters. After pixel-wise aggregation, we use hierarchical residual-like connections to improve the multi-scale representation ability at a more granular level. This block can be plugged into the state-of-the-art backbone without any effort. Figure 4 shows the design of the DDR module.

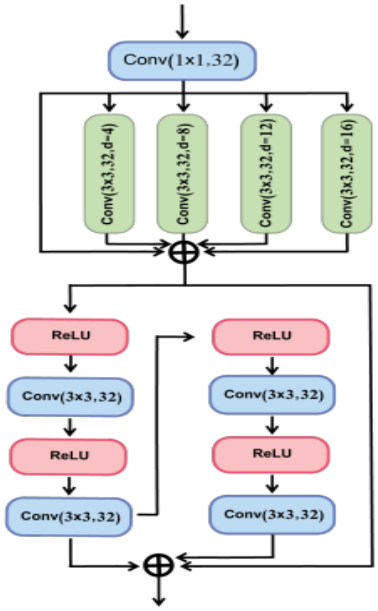


Figure 4. Visual Representation of dilated residual convolution module

3.2.3. Convolutional Block Attention Module (Cbam)

We use a lightweight spatial and channel attention module after the dilated residual convolution block to focus on the relevant features while diminishing the other parts [26]. The spatial attention extracts the inter-spatial relationships of features to find "where"' is an informative part of the image. To calculate this, we first apply average pooling and max pooling, which summarize the average presence of features and the most activated presence of a feature, respectively. Then, we use a convolution layer in addition to the concatenated feature descriptor to create a spatial attention map that specifies where to highlight or suppress features. [26].
The channel attention block redistributes the channel's feature responses to give higher importance to specific channels over others. In order to compute the channel attention, we squeeze the spatial dimension of the input feature map. [26].

3.3. Loss Function

In an image, the edge and non-edge pixel data are not equally distributed. CNN models can achieve pretty high accuracy just by predicting the majority class, but they fail to capture the minority class. Unfortunately, this accuracy is misleading. To address this problem, we adopt the weighted cross-entropy loss function proposed in [19].

To train the network, we match all the stages and fused outputs to the ground truth. The following equation compares each pixel of each image to its label as.

$$L(x_i;W)=\begin{cases}\alpha.log(1-P(x_i;W)) & if y_i=0\\ 0 & if \le y_i \le \eta\\ \beta.logP(x_i;W) & otherwise,\end{cases} \quad (1)$$

$$\alpha=\lambda.\frac{|Y^+|}{|Y^+|+|Y^-|} \qquad \beta=\frac{|Y^-|}{|Y^+|+|Y^-|} \quad (2)$$

X, P(X), Y, W, and η, respectively, denote features extracted from the CNN network, the output of the standard sigmoid function, the ground truth edge probability, all the parameters that will be learned in the CNN network, and the percentage of non-edge and edge pixels. The hyper-parameter is used to balance the number of positive and negative samples. Because each image is being labelled by multiple annotators, and humans vary in cognition, the predefined threshold is used to distinguish between edge and non-edge pixels in the edge probability map. If a pixel is marked by fewer than η of the annotators, then it is considered a non-edge pixel. To generalize the loss function to all the pixels inside the image (I), at each stage (k) and fuse layer, the following loss function is used:

$$L(W) = \sum_{i=1}^{|I|} \left(\sum_{k=1}^{|K|} L\left(x_i^k; W\right) + L\left(x_i^{fuse}; W\right) \right) \qquad (3)$$

4. EXPERIMENTS AND DISCUSSIONS

4.1.1. Implementation Details

We use PyTorch for implementation and initialize the stages of our backbone networks with Gaussian distribution with zero-mean and standard deviation of 0.01. The learning rate starts from 0.01 and then is updated using a linear scaling factor, multiplying 0.1 for every two epochs. The optimizer is stochastic gradient descent, and the training process terminates at eight epochs. We conduct all the experiments on a single GPU, NVIDIA GeForce 2080Ti, with 11G memory.

4.1.2. Dataset

In order to have a fair comparison to other published works in tables 1 and 2, we evaluate our proposed network on the same Berkeley Segmentation (BSDS500) [8] and NYUDv2 [9] Dataset . BSDS500 consists of 200 training, 100 validation, and 200 test images. We combine the 200 training images with 100 validation images to create a training set. We adopt the data augmentation technique similar to RCF [19]. In addition, similar to RCF, we also added the PASCAL VOC [27] dataset and its flipped images into our training set.
The NYUD dataset is composed of 1449 densely labelled pairs of aligned RGB and depth images (HHA). This dataset consists of video sequences from various indoor scenes captured by the Microsoft Kinect's RGB and Depth cameras. It is divided into 381 training, 414 validation, and 654 testing images. Similar to RCF [19], we rotate the images and corresponding annotations to 4 different angles (0, 90, 180, and 270 degrees) and flip them at each angle.

4.1.3. Performance Metrics

Note that the share of edge pixels in each image is around 10%, whereas the share of non-edge pixels is 90%. Therefore, even when a model fails to predict any edges, its accuracy is still 90%. As such, accuracy is a poor measure for evaluating imbalanced problems such as edge detection. Therefore, we use F-Score for the evaluation of our model. The F-score combines the precision and recall of the model, where it reaches its best value at one and its worst score at 0.

Recall = TruePositives / (TruePositives + FalseNega- tives)

Precision= TruePositives / (TruePositives + FalsePosi- tives)

F-Measure = (2 * Precision * Recall) / (Precision + Recall) and

We need a threshold to binarize the output of the CNN network to make it comparable to the ground truth, which is also binarized. There are two ways to compute the optimal threshold corresponding to the F-score.

- Optimal Dataset Scale: Iterates over all possible thresholds and sets one threshold for the entire dataset. The threshold that gives the best F-score for the dataset is used to calculate ODS score.
- Optimal Image Scale: Finds the best threshold and corresponding F-score for each image. The OIS F-score is calculated by averaging all of the F-scores for all images.

4.1.4. Comparison with State-of-the-Arts

On the BSDS500 dataset: We compare our methods in terms of F-score and number of parameters to prior edge detection approaches, including both traditional ones and recently proposed CNN-based models. According to Table 1 and Figure 5, we notice that our baseline model, while using a significantly lower number of parameters, can even achieve outstanding results (ODS of 0.792 and OIS of 0.805) which are equal or better than most recent lightweight CNN models such as BDCN2, TIN1,TIN2, FINED3-Inf and FINED3-Train [22].

Table 1. Comparison to other methods on BSDS500 dataset.

Method	ODS	OIS	#P (million)
Canny	0.611	0.676	-
OEF	0.746	0.77	-
gPb-UCM	0.72	0.755	-
SE	0.743	0.763	-
AMHNET	0.798	0.829	22
BDP-Net	0.808	0.828	18.7
FCL-Net	0.826	845	16.5 M
BAN	0.81	0.827	15.6
LPCB	0.815	0.834	15.7
BMRN	0.828	0.81	+14.8
RCF	0.806	0.823	14.8
HED	0.788	0.808	14.7
COB	0.793	0.82	28.8
RHN	0.817	0.833	11.5
CED	0.815	0.834	21.4
DeepEdge	0.753	0.772	-
DeepContour	0.757	0.776	0.38
BDCN	0.82	0.838	16.3
BDCN2	0.766	0.787	0.48
BDCN3	0.796	0.817	2.26
BDCN4	0.812	0.83	8.69
TIN1	0.749	0.772	0.08
TIN2	0.772	0.792	0.24
FINED3-Inf	0.788	0.804	1.08
FINED3-Train	0.79	0.808	1.43
Our model	0.792	0.805	0.506

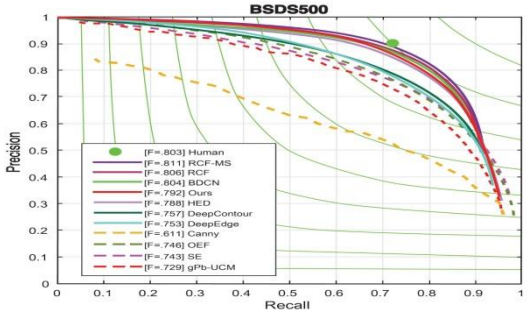


Figure 5. Precision-Recall curves of our models and some competitors on BSDS500 dataset

.

On the NYUD dataset: The comparison results on the NYUD dataset are illustrated in Table 2, and the precision- recall curves are depicted in Figure 6. For testing the model on NYUD, we use network settings similar to that used for BSDS500. Some studies use two separate models to train RGB images and HHA feature images of NYUD and report the evaluation metrics on the average for the outputs of the models. Our network is only tested on RGB images, so in order to evaluate results fairly, we contrasted our model's output with those of models that were only tested on RGB.

Table 2. Comparison with other methods on NYUD dataset.

Method	ODS	OIS	#P (million)
OEF	0.651	0.667	_
gPb-UCM	0.632	0.661	_
SE	0.695	0.708	_
SE+NG+	0.706	0.734	_
AMHNET	0.744	0.758	22
BDCN	0.748	0.763	16.3
LPCB	0.739	0.754	15.7
RCF	0.743	0.757	14.8
BMRN	0.759	0.776	+14.8
HED	0.72	0.734	14.7
Our Model	0.725	0.738	0.5

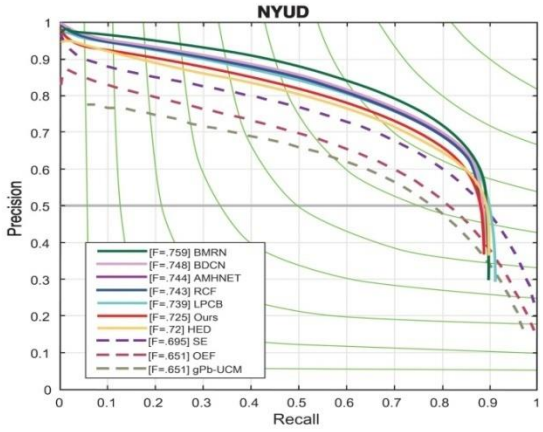


Figure 6. Precision-Recall curves of our models and some competitors on the NYUD dataset.

5. CONCLUSIONS

Edge detection has numerous practical applications in the real world; hence, we must design an efficient architecture for its implementation. Most existing deep neural networks for edge detection tasks use transfer learning from pre-trained models such as VGG16, which have a large number of parameters and are trained for high-level tasks. However, edge detection has a simple set of features and does not require a large number of convolutional layers for feature extraction. Therefore, in this research, we introduced a new architecture that is both lightweight and has state-of-the-art performance. Our network makes full use of customized depthwise separable and deformable convolutions to carry out edge detection. Besides, we use lightweight components to increase the receptive field of our model to produce high-quality edges. Our network architecture is extendable and can potentially be employed for use in other vision tasks such as salient object detection and semantic segmentation.

REFERENCES

[1] Victor Wiley and Thomas Lucas. "Computer vision and image processing: a paper review. International Journal of Artificial Intelligence Research", 2(1):29–36, 2018.

[2] Ronald J Holyer and Sarah H Peckinpaugh. Edge detection applied to satellite imagery of the oceans. IEEE transactions on geoscience and remote sensing, 27(1):46–56, 1989.

[3] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array, 10:100057, 2021.

[4] Wei-Chun Lin and Jing-Wein Wang. Edge detection in medical images with quasi high-pass filter based on local statistics. Biomedical Signal Processing and Control, 39:294–302, 2018.

[5] Jan Kristanto Wibisono and Hsueh-Ming Hang. Traditional method inspired deep neural network for edge detection. In 2020 IEEE International Conference on Image Processing (ICIP), pages 678–682. IEEE, 2020.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[7] Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8368–8375, 2019.

[8] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence, 33(5):898–916, 2010.

[9] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European conference on computer vision, pages 746–760. Springer, 2012.

[10] O Rebecca Vincent, Olusegun Folorunso, et al. A descriptive algorithm for sobel image edge detection. In Proceedings of informing science & IT education conference (InSITE), volume 40, pages 97–107, 2009.

[11] Renjie Song, Ziqi Zhang, and Haiyang Liu. Edge connection based canny edge detection algorithm. Pattern Recognition and Image Analysis, 27(4):740–747, 2017.

[12] Rajiv Mehrotra and Shiming Zhan. A computational approach to zero-crossing-based two-dimensional edge detection. Graphical Models and Image Processing, 58(1):1–17, 1996.

[13] James J. Clark. Authenticating edges produced by zero-crossing algorithms. IEEE Trans- actions on Pattern Analysis and Machine Intelligence, 11(1):43–57, 1989.

[14] Scott Konishi, Alan L. Yuille, James M. Coughlan, and Song Chun Zhu. Statistical edge detection: Learning and evaluating edge cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(1):57–74, 2003.

[15] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. IEEE transactions on pattern analysis and machine intelligence, 37(8):1558–1570, 2014.

[16] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision, pages 1395–1403, 2015.

[17] Yupei Wang, Xin Zhao, and Kaiqi Huang. Deep crisp boundaries. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3892–3900, 2017.
[18] Dan Xu, Wanli Ouyang, Xavier Alameda-Pineda, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. Advances in neural information processing systems, 30, 2017.
[19] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3000–3009, 2017.
[20] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In Proceedings of the European Conference on Computer Vision (ECCV), pages 562–578, 2018.
[21] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. In Proceedings of the IEEE/CVF Confer- ence on Computer Vision and Pattern Recognition, pages 3828–3837, 2019.
[22] Jan Kristanto Wibisono and Hsueh-Ming Hang. Fined: Fast inference network for edge detection. arXiv preprint arXiv:2012.08392, 2020.
[23] Xavier Soria Poma, Edgar Riba, and Angel Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1923–1932, 2020.
[24] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017.
[25] Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Mar- tin Reuter. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. Neu- roImage, 219:117012, 2020.
[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.
[27] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fi- dler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 891–898, 2014.

A MACHINE LEARNING/DEEP LEARNING HYBRID FOR AUGMENTING TEACHER-LED ONLINE DANCE EDUCATION

Catherine Hung

Palo Alto Senior High School, Palo Alto, USA

ABSTRACT

For online dancers, learning a dance move properly without the feedback of a live instructor can be challenging because it is difficult to determine whether a move is done correctly. The lack of proper guidance can result in doing a move incorrectly, causing injury. In this work – we explore the use of a hybrid Deep Learning/Machine Learning approach to classify dance moves as structurally correct or incorrect. Given a video clip of the dancer doing a move, such as the grand plie, the algorithm should detect the correctness of the movement. To capture the overall movement, we proposed various methods to process data, starting with deep learning techniques to convert video frames into landmarks. Next, we investigate several approaches to combining landmarks from multiple frames and training machine learning algorithms on the dataset. The distinction between correct and incorrect grand plies achieved accuracies of over 98%.

KEYWORDS

Deep Learning, Machine Learning, Classification, Online Dance Education

1. INTRODUCTION

Dancing attracts many people due to its numerous health benefits ranging from enhancing cognitive function to improving balance [1]. Additionally, many individuals dance to improve their mood and overall well-being [2]. In 2021, around 24.71 million people in the US, or around 7.5% of the US population, took advantage of the benefits of dance [3].

While many attend in-person dance classes, due to the COVID-19 pandemic and lockdown in 2020, dance studios were required to stop in-person learning. Thus, many resorted to online dance instruction and dance apps. STEEZY, which has over one million downloads, is an app that teaches users how to dance in various styles ranging from K-pop to ballet for novices, intermediate, and advanced dancers [4]. Even after the lockdown, many still attend online dance classes because they allow for more flexibility in one's schedule, are typically cheaper than in-person classes, and provide people in remote areas the opportunity to participate in classes and learn from teachers all around the world [5]. Yet, despite these advantages, without a teacher physically present to correct them, dancers may incorrectly learn dance moves. Also, it can be hard for dance teachers to correct many students online without some physical contact, and it is difficult for inexperienced new dancers to self-correct [6].

Incorrectly doing even a basic dance move by using incorrect technique and poorly aligning one's body can injure dancers, preventing them from furthering their dance education and learning more complicated moves [7].

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 29-40, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130403

Ballet is a classical style of dance whose movements and technique serve as the foundation of other dance styles [8]. In ballet, a grand plie, an elementary move, helps develop the balance and stability needed for other moves [9]. For this reason, it is important to be able to do this move correctly. It involves the bending of the knees up until the thighs are horizontal to the floor while maintaining a straight back. While this move seems straightforward in theory, there are several ways in which dancers can incorrectly perform this move, such as by incorrectly bending one's spine or abruptly moving from one pose to another after a long pause. This can result in injuries such as muscle strain [10]. Because of this, it is critical for online dancers to learn how to perform each dance move correctly.

AI has the potential to detect and differentiate incorrect dance moves from correct dance moves given its demonstrated effectiveness across a range of sectors including the detection of correct mask-wearing and sports [11]. This paper explores how deep learning and machine learning techniques can be leveraged in the dance teaching domain. Dance teachers can use the feedback on the correctness of a move to determine areas of growth for students.

To build an AI that accurately detects correct and incorrect grand plies, we gathered videos of dancers performing this move correctly and incorrectly and consulted an experienced dancer to label the videos for our dataset. We then used a deep learning technique to extract the physical landmarks of each video frame. Subsequently, we processed these landmarks in a variety of ways to assess different aspects of the motion, such as speed and smoothness. After that, we used our processed landmarks to train three well-known and robust machine learning algorithms: K-Nearest Neighbors, Random Forest Classifier, and MLP. We then evaluated the results.

After using different techniques to process our data and train our models, we were eventually able to achieve accuracies over 98% and occasionally 100%. Overall, we achieved similar results for all three of the machine learning algorithms we ran our processed data on.

Our paper outline is as follows. Section 2 discusses related work and our novel contributions. Section 3 examines how our solution works, Section 4 covers our results and analysis, Section 5 discusses our observations, and Section 6 notes our conclusions and future work.

2. RELATED WORK

Improving dance education is a field of interest to many. Multiple studies have looked into ways to implement technology into correcting dance movements. For example, past studies have used physical sensors that are attached to dancers to track the movements of their body parts [5,12]. Other studies have collected data through cameras and analyzed the data with machine learning frameworks, such as OpenPose [13,14,15].

Similarly, some studies are focused on detecting the movement of multiple dancers at a time to possibly be utilized in a class with multiple dancers [16,14]. Other studies are focused on tracking the movement of a single dancer or athlete [12, 13, 17]. For example, Woah.AI is an app that teaches users how to do modern TikTok dances and uses AI to provide feedback [13].

Having data on the body movements of dancers can be used to benefit different aspects of dance. One study is using this data to analyze how dance-related injuries develop over time in order to find methods to prevent them [12]. Whereas, another study is concerned with generating dances that are more realistic, creative, and appealing based on specific guidelines such as composition, performance, and evaluation [18].

Our study differs from the related work above in several ways. First, our goal is to detect and correct dance moves. This is unlike other research that is intended to generate new dances or track how dance injuries develop over time. Additionally, unlike how many researchers are detecting dance moves through physical sensors, which may be uncomfortable to dance with, we are using non-invasive techniques such as cameras to detect dance moves. We expect our approaches to be more practical for use in dance education. Also, other researchers are looking at how to detect the dance moves of multiple dancers in a room. We, however, are only detecting the dance moves of a single dancer since our research is aimed toward online dancers.

Online dance classes are challenging for both students and teachers. One study that analyzed virtual dance education found that while it is progressing promisingly, there is still improvement needed to be done [19]. In online streaming platforms like Zoom, because students are represented by tiny squares, the teacher is unable to see an individual student in detail and has difficulty tracking the issues of multiple students at the same time. Our application has the potential to overcome these limitations since each video feed can be analyzed separately.

3. SOLUTIONS

We begin by outlining the requirements for an effective solution.

3.1. Requirements

A. Correct movement requires the integration of the full body. Therefore, in order to properly assess the movement, the algorithm should analyze a full-body video clip of the dancer doing a grand plie.

B. Given the full-body video clip, the algorithm should be able to accurately detect whether the dancer is correctly or incorrectly doing the grand plie.

C. Smoothness and continuity are important elements of a correct grand plie [10]. Therefore, the algorithm should be able to leverage the data to assess the smoothness and continuity of a given grand plie.

D. In addition to smoothness, the algorithm should also be able to detect other essential factors when classifying the grand plie as correct or incorrect. One important factor is the placement of the hips. In a correct grand plie, the hips are between and in line with the knees. Whereas, in an incorrect grand plie, dancers place their hips behind or below their knees. Another factor is the placement of the heels. Correct grand plies involve the heels naturally and slightly lifting up as dancers are bending their knees. However, dancers who incorrectly do grand plies forcibly push their heels up high [10].

3.2. How Our Solution Works

Next, we outline our algorithmic approach and our dataset.

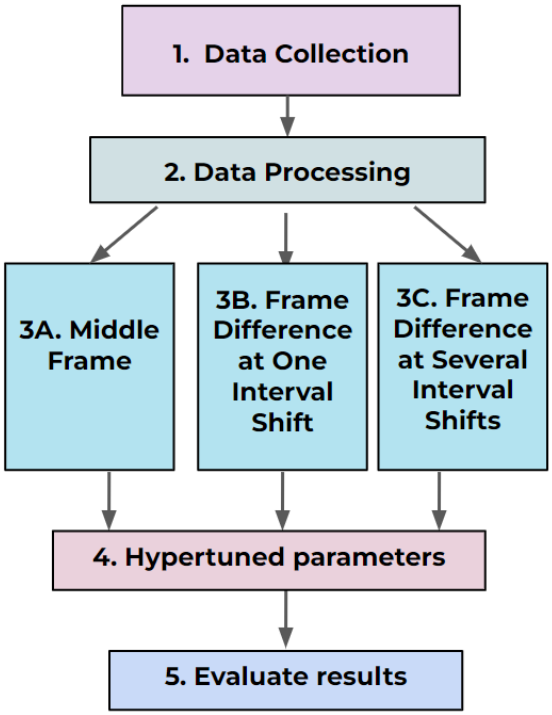


Figure 1. Experimental Process

Figure 1 describes our overall process in creating our model that detects whether or not a dancer is doing a grand plie correctly. For our first step, we gathered videos of correct and incorrect grand plies. Then, in Step 2, we processed these videos using deep learning algorithms to generate landmarks. Next, in Steps 3A, 3B, and 3C, we combined the landmarks from different frames using three different methods. Then, in Step 4, we trained our data on machine learning algorithms. Finally, we analyzed our findings in Step 5.

3.2.1. Data

In total, we collected 42 video clips of dancers correctly doing a grand plie and 34 video clips of dancers incorrectly doing a grand plie. These videos came from various YouTube videos, Instagram posts, and live dancers. These video clips were labeled by a trained dancer and we had permission to use them.

3.2.2. Data Processing

To convert the video clips into features, we used MediaPipe Pose, a machine learning solution for tracking the locations of body parts [20]. Given a full-body RGB multi-frame video of a person, MediaPipe Pose uses BlazePose, a built-in neural network solution, to locate the person of interest and identify the locations of the 33 landmarks, such as the left leg, right hip, and left ankle [21]. Within each landmark, the x-component, y-component, the landmark depth (z), and the likelihood of a landmark being visible in the image (v) are given as numerical values. Thus, for each video frame, 132 data points are collected. The label of the video is one of two values — whether the move is correctly or incorrectly executed. Additionally, using MediaPipe Pose also allows our approach to be deployed in a variety of delivery vehicles, such as in a mobile app or website.

Figure 2. MediaPipe Pose for tracking body position. The image was provided by a volunteer who consented to its use.

3.2.3. Processing the MediaPipe Pose Output

We experimented with processing the MediaPipe Pose output in three ways – listed as 3(A-C) below. The specific calculations are listed in the equations section.

3A. Initially, we used one frame per video for our dataset. We theorized that with a captured moment in the middle frame of each video clip, the AI algorithm could detect flaws in body placement, such as the intentional lifting of the heels. We define the middle frame as the frame in the exact middle of the video. The middle of the grand plie is when dancers' errors such as the sticking out of hips and the intentional lifting of the heels are the most prominent [10].

While our first method could potentially capture snapshots of error, such as the intentional lifting of the heels, it fails to consider the overall smoothness of the movement. When incorrectly doing the grand plie, dancers tend to pause for a while when their legs are bent the most, whereas correct grand plies involve gradual, nonstop movement [10]. To capture this movement, we took the difference in the landmark coordinates between two specified frames at certain distances from the middle frame of the video. If these two frames have some difference, the dancer could be gradually moving throughout the grand plie. Little to no change in the difference between these two frames indicates that the dancer is pausing at a pose. Furthermore, a large difference shows that the dancer could be abruptly moving from one move to another. This approach is illustrated in 3B and 3C.

3B. We initially calculated the difference in landmark positions between two frames, one frame that was twenty frames before the midpoint of the video and another frame that was twenty frames after the video midpoint. Our new dataset is comprised of taking the frame difference at one interval shift per video.

3C. To increase the size of our dataset, we calculated the landmark difference of two frames at a total of three different interval shifts. We define an interval shift as a time location or frame

shifted by an interval relative to the middle of the video. We extended method 3B by adding two additional interval shifts. We added frame differences between a frame that was 15 frames before the midpoint and a frame that was 25 frames after the midpoint. We also calculated the difference between a frame 25 frames before the midpoint and another one 15 frames after the midpoint. Our dataset for 3C thus consisted of triple the number of samples in 3B.

3.3. Equations

We name the equation parameters as below.

Videos (*1..N*)
Frames - *F. F_i (frames in video i)*
Frame midpoint = $M_i = Floor(F_i/2)$
Left_bookend = *Lb*
Right bookend = *Rb*
Length $L = Rb - Lb$

3.3.1. Middle Frame (3A)

$$M_i \tag{1}$$

Each row has 132 features for each frame in the video. We use the frame *Mi* that is in the middle of each video. What this means is that the first group of data has *N* samples. Each sample represents one video which contains the middle frame of each video.

3.3.2. Frame Difference at One Interval Shift (3B)

We used Length = 40. The Left bookend Lb and Right bookend Rb are as follows.

$$Lb = M_i - 20 \tag{2}$$
$$Rb = M_i + 20 \tag{3}$$

We convert each feature to the difference between that feature's value in Frame *Rb* and the value in Frame *Lb*.

For example:

$$\text{Nose\_x} = \text{Frame } (Rb)\text{\_Nose\_x} - \text{Frame } (Lb)\text{\_Nose\_x} \tag{4}$$

What this means is that the second group of data has *N* samples. Each sample represents one video. Each feature is the difference of that feature's value between Frames *Lb* and *Rb* where the sample has Left and Right bookends as calculated in equations (2) and (3).

3.3.3. Frame Difference at Several Interval Shifts (3C)

We used Length = 40. The Left bookend Lb and Right bookend Rb are as follows.

$$Lb = M_i - 25 \tag{5}$$
$$Rb = M_i + 15 \tag{6}$$
$$Lb = M_i - 25 \tag{7}$$
$$Rb = M_i + 15 \tag{8}$$

What this means is that the third group of data has *3N* samples. Each video produces three different samples. Each feature is the difference of that feature's value between Frames *Lb* and *Rb* where the first sample has (2) and (3), the second sample has (5) and (6), and the third sample has (7) and (8).

In response to the unbalanced dataset with more correct grand plies, we utilized SMOTE to oversample incorrect grand plies [22]. SMOTE, also known as Synthetic Minority Oversampling Technique, aims to create a balanced data set by replicating some of the data from the minority class to make the size of the minority class the same as that of the majority class.

3.4. Final Algorithm

We used three different ML algorithms — K-Nearest Neighbors, Random Forest Classifier, and MLP — to test our model and hypertuned parameters within each algorithm to obtain the best accuracy.

- Random Forest Classifier is a classification method composed of many decision trees. Each decision tree, which is built through random methods, outputs a class prediction and the Random Forest Classifier outputs the most popular class prediction [23]. With Random Forest Classifier, the data iterated through an interval of the number of estimators ranging from 1 to 13 and through maximum depths ranging from 10 to 40.
- K-Nearest Neighbors, also known as KNN, is another classification method that presumes that similar data points with the same class are located close together. To classify a data point, it looks at the classes of the other close data points, or neighbors, and determines the most popular class. With KNN, the data iterated through an interval of the number of neighbors ranging from 1 to 10 neighbors [24].
- Multilayer Perceptron, or MLP, is a basic neural network that consists of multiple layers including an input layer, several hidden layers, and an output layer [25]. With MLP, the training iterated through different learning rates ranging from 0.01 to 0.1 and maximum epochs ranging from 20 to 125. Early stopping was enabled and the tolerance was set to 0.00001.

4. RESULTS

Overall, throughout the three different methods of processing data, algorithms with and without SMOTE resulted in similar maximum accuracies. In the cases with a difference in accuracies, using an algorithm with SMOTE slightly lowered the highest accuracy. For example, this was the case in Fig. 3; MLP had an accuracy of 0.7826 without SMOTE and an accuracy of 0.7391 with SMOTE. Given that there were more correct samples than incorrect samples, the algorithms could have been biased to predict that a given video is correct without SMOTE. Additionally, all three algorithms had the highest accuracy when given a dataset with frame differences at several interval shifts per video clip. In Fig. 5, all maximum accuracies were above 90% with Random Forest Classifier even achieving an accuracy of 100%. The three algorithms had the lowest accuracy when a frame difference was only calculated at one interval shift per video. Fig. 4 demonstrates that the single frame difference approach (3B) did not generate good accuracies.

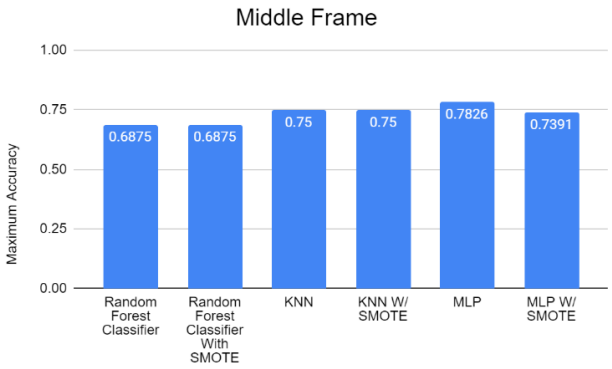


Figure 3. Middle Frame Method (3A)

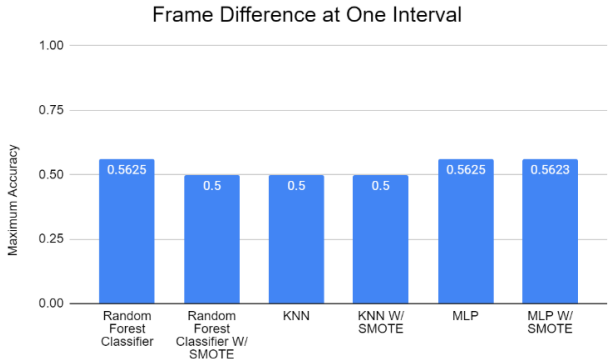


Figure 4. Frame Difference at One Interval Method (3B)

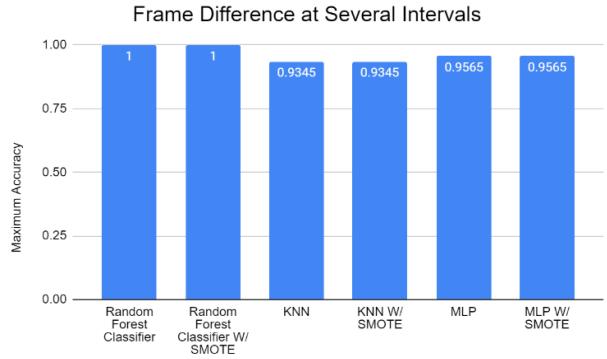


Figure 5. Frame Difference at Several Intervals Method (3C)

Overall, these three algorithms performed roughly the same, and there was not an algorithm that reached a maximum accuracy that was drastically different from the other algorithms. In Fig. 5, Random Forest Classifier performed slightly higher for frame differences at several interval shifts, achieving a maximum accuracy of 100%.

The tables below display our results. Each table consists of the hyperparameters for each algorithm as well as the highest and lowest accuracy that resulted with hyperparameter tuning. There is a table for each of our 3 methods: 3A, 3B, and 3C.

Table 1. Results of Using Middle Frame Method (3A)

Algorithms	**Hyperparameters**	**Accuracy**
Random Forest Classifier	Estimators (1-13), Max Depth (10-40)	0.5625-0.6875
KNN	#Neighbors (1-10)	0.5625-0.75
MLP	Learning rate (0.01-0.1), Max Iteration (20 - 125)	0.3478 - 0.7826

Based on Table 1, while MLP achieved the highest accuracy (0.7826) than the other two algorithms, the lowest accuracy was 0.3478. Whereas, even though the highest accuracies from KNN and Random Forest Classifier are lower than that of MLP, they both had the lowest accuracies of 0.5625. Therefore, MLP required more tuning to get to a higher accuracy than the other two algorithms.

Table 2. Results of Using Frame Difference at One Interval Method (3B)

Algorithms	**Hyperparameters**	**Accuracy**
Random Forest Classifier	Estimators (1-13), Max Depth (10-40)	0.3125-0.6875
KNN	#Neighbors (1-10)	0.125-0.5
MLP	Learning rate (0.01-0.1), Max Iteration (20 - 125)	0.3158 - 0.5625

Referring to Table 2, Random Forest Classifier and KNN both had the same difference between the highest and lowest accuracy (0.375), whereas MLP had a slightly smaller range (0.2467). Therefore, for this solution, MLP required less tuning than Random Forest Classifier and KNN.

Table 3. Results of Using Frame Difference at Several Intervals Method (3C)

Algorithms	**Hyperparameters**	**Accuracy**
Random Forest Classifier	Estimators (1-13), Max Depth (10-40)	0.7826 - 1
KNN	#Neighbors (1-10)	0.4783 - 0.9348
MLP	Learning rate (0.01-0.1), Max Iteration (20 - 125)	0.4783 - 0.9565

Referring to Table 3, Random Forest Classifier had the lowest accuracy range (0.2174) compared to KNN and MLP, which had accuracy ranges of both 0.4565 and 0.4782 respectively. These ranges are both more than double that of Random Forest Classifier. Therefore, for this solution, Random Forest Classifier required the least amount of tuning.

5. DISCUSSION

The three different algorithms – KNN, MLP, and Random Forest Classifier – all had around the same accuracy for each method of data processing. For all three algorithms, taking the frame difference at several different interval shifts of each video resulted in the highest accuracy, which was above 90% according to Fig. 4. Whereas, taking the frame difference at only one interval of each video resulted in the lowest accuracy (50 - 68.75%) for each of the algorithms which can be seen in Fig. 3.

Having the frame difference at three different interval shifts per video tripled the sample size to 228 and also increased the test size to around 46. There could have been a higher accuracy because there was more training data (around 182) to practice on and more samples to test on. Having more frame shifts at different time locations of the movement provides more insight on correct versus incorrect grand plies. In Fig. 4, the algorithms performed the same with and without SMOTE. Even though there were 24 more sample videos of correct grand plies than incorrect grand plies, the sample sizes for each, both over 100, could have been large enough for the algorithm to not be favored toward correct grand plies.

A limitation of the study was the small sample size. There were only 76 total video clips – 42 of dancers correctly doing a grand plie and 34 of dancers incorrectly doing grand plies. Even though a difference of 8 samples would not be much for a larger dataset, for a smaller sample size like 76, this difference could be impactful enough to introduce bias favored toward correct grand plies, which could have been the case in Fig. 2 and Fig. 3. For these charts, Random Forest Classifier and KNN performed slightly worse with SMOTE where the number of correct and incorrect samples is the same. Additionally, due to a small sample size, the test set, which was 20% of the sample size or around 15, was also small, which did not give as much variation in the accuracies. However, increasing the test set would simultaneously decrease the training set, so the algorithm may not have many samples.

6. CONCLUSIONS AND FUTURE WORK

To allow more online dancers to be successful and to combat the possible injuries that could arise with online dance classes, we built an AI system that detects correct and incorrect grand plies to a high level of accuracy.

SMOTE was used to account for the imbalance of correct and incorrect videos. We then ran it through three algorithms – Random Forest Classifier, KNN, and MLP — and hypertuned parameters. Taking the frame difference of frames at multiple interval shifts was the most successful. Capturing the position of the dancer at three different time locations possibly increased the robustness of the AI system performance. The three different algorithms performed similarly in terms of accuracy for the three different methods.

In the future, it would be interesting to combine our method of capturing the frame intervals with capturing the middle frame of the videos to determine if the results differ. Further work could also be done to identify the ideal frame intervals and the number of frame shifts for each dance move.

This method could also be tested to detect the correctness of more complicated dance moves, such as a turn, a leap, or even a series of dance moves, such as a grand plie and a turn in one video clip. Furthermore, given that MediaPipe Pose permits this, our solution could be deployed as a mobile app or a website that could be compatible with Zoom and other virtual school platforms.

REFERENCES

[1] M. Alricsson, K. Harms-Ringdahl, K. Eriksson, and S. Werner, "The effect of dance Training on Joint Mobility, Muscle Flexibility, Speed and Agility in Young Cross-country Skiers - a Prospective Controlled Intervention Study." *Scandinavian Journal of Medicine & Science in Sports*, vol. 13, no. 4, 14 July 2003, pp. 237-43. *National Library of Medicine*, https://doi.org/10.1034/j.1600-0838.2003.00309.x. Accessed 6 July 2022.

[2] A. Maraz, O. Király, R. Urbán, M. D. Griffiths, and Z. Demetrovics, "Why Do You Dance? Development of the Dance Motivation Inventory (DMI)." *PLOS ONE*, vol. 10, no. 3, 24 Mar. 2015, p. e0122866. *National Library of Medicine*, https://doi.org/10.1371/journal.pone.0122866. Accessed 6 July 2022.

[3] D. Lange. "Dance, step, and other choreographed exercise participants US 2021." *Statista*, February 2022, https://www.statista.com/statistics/756629/dance-step-and-other-choreographed-exercise-participants-us/. Accessed 6 July 2022.

[4] "Steezy Studio: Reach Your Dance Goals." *STEEZY Studio | Reach Your Dance Goals*, https://www.steezy.co/#levels.

[5] "Pros and cons of doing dance classes online - go&dance." *Go&Dance*, https://www.goandance.com/en/blog/lifestyle/72-pros-and-cons-of-doing-dance-classes-online. Accessed 6 July 2022.

[6] Bellerose, Samantha. "Should you enroll your child in online virtual dance lessons?" *Dance Parent 101*, https://danceparent101.com/should-you-enroll-your-child-in-online-virtual-dance-lessons/. Accessed 6 July 2022.

[7] A. Malkogeorgos, F. Mavrovouniotis, G. Zaggelidis, and C. Ciucurel, "Common dance related musculoskeletal injuries." *Journal of Physical Education and Sport. Research Gate*, www.researchgate.net/publication/287785890_Common_dance_related_musculoskeletal_injuries. Accessed 6 July 2022.

[8] H. Kantilaftis. "Ballet and modern dance: using ballet as the basis for other dance techniques." *New York Film Academy*, 5 August 2014, https://www.nyfa.edu/student-resources/ballet-and-modern-dance/. Accessed 6 July 2022.

[9] T. Tellier. "Plié power." *Everyday Ballet*, 10 June 2017, https://www.everydayballet.com/plie-power/. Accessed 6 July 2022.

[10] I. Levendosky. "The ups and downs of demi and grand plies." *Front Range Classical Ballet Academy*, 14 October 2018, https://frcballet.com/blog/2018/10/11/the-ups-and-downs-of-demi-and-grand-plies. Accessed 6 July 2022.

[11] J. Tomás, A. Rego, S. Viciano-Tudela, and J. Lloret. "Incorrect facemask-wearing detection using convolutional neural networks with transfer learning." *Healthcare*, vol. 9, no. 8, 16 Aug. 2021, p. 1050, https://doi.org/10.3390/healthcare9081050.

[12] UNT Health Center. "Using 3D motion technology to prevent injuries in professional dancers." *The DO*, 8 October 2018, https://thedo.osteopathic.org/2018/10/using-3d-motion-technology-to-prevent-injuries-in-professional-dancers/. Accessed 6 July 2022.

[13] "Woah! You Can Really Dance!" *Woah*, https://www.woah.ai/.

[14] M. Trajkova. "Designing AI-Based Feedback for Ballet Learning." *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020): n. pag.

[15] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." *Arxiv. Cornell University*, https://doi.org/10.48550/arXiv.1812.08008. Accessed 19 Jan. 2023.

[16] A. Faridee, S. R. Ramamurthy, H. M. S. Hossain, and N. Roy, "HappyFeet: recognizing and assessing dance on the floor." *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications* (2018): n. pag.

[17] A. Luna, et al. "Artificial Intelligence application versus physical therapist for squat evaluation: a randomized controlled trial." *Scientific Reports*, vol. 11, no. 1, 13 Sept. 2021, https://doi.org/10.1038/s41598-021-97343-y.

[18] S. Ibrahimi. "Composition, performance and evaluation: a dance education framework for AI systems." International Conference on Computational Creativity 2021, 2021.

[19] You, Yuhui. "Online Technologies in Dance Education (China and Worldwide Experience)." *Research in Dance Education*, 14 Oct. 2020, pp. 1-17. *Taylor and Francis Online*, https://doi.org/10.1080/14647893.2020.1832979. Accessed 9 Sept. 2022.

[20] "Pose-Mediapipe."*Google*, https://google.github.io/mediapipe/solutions/pose.html. Accessed 6 July 2022.

[21] V. Bazarevsky, et al. "BlazePose: on-device real-time body pose tracking." *arXiv e-prints*, https://doi.org/10.48550/arXiv.2006.10204. Accessed 6 July 2022.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic Minority Over-sampling Technique*. ArXiv, 2011. *Arxiv*, https://doi.org/10.48550/arXiv.1106.1813.

[23] "sklearn.ensemble.RandomForestClassifier—scikit-learn 1.1.1 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed 7 July 2022.
[24] "sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.1.1 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. Accessed 7 July 2022.
[25] "sklearn.neural_network.MLPClassifier — scikit-learn 1.1.1 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Accessed 7 July 2022

AUTHORS

Catherine Hung is a senior at Palo Alto High School in Palo Alto, California. Her research interests are artificial intelligence and machine learning.

MACHINE LEARNING CHATBOT FOR SENTIMENT ANALYSIS OF COVID-19 TWEETS

Suha Khalil Assayed, Khaled Shaalan, Manar Alkhatib, Safwan Maghaydah

Faculty of Engineering and IT, The British University in Dubai, UAE

ABSTRACT

The various types of social media were increased rapidly, as people's need to share knowledge between others. In fact, there are various types of social media apps and platforms such as Facebook, Twitter, Reddit, Instagram, and others. Twitter remains one of the most popular social application that people use for sharing their emotional states. However, this has increased particularly during the COVID-19 pandemic. In this paper, we proposed a chatbot for evaluating the sentiment analysis by using machine learning algorithms. The authors used a dataset of tweets from Kaggle's website, and that includes 41157 tweets that are related to the COVID-19. These tweets were classified and labelled to four categories: Extremely positive, positive, neutral, negative, and extremely negative. In this study, we applied Machine Learning algorithms, Support Vector Machines (SVM), and the Naïve Bayes (NB) algorithms and accordingly, we compared the accuracy between them. In addition to that, the classifiers were evaluated and compared after changing the test split ratio. The result shows that the accuracy performance of SVM algorithm is better than Naïve Bayes algorithm, even though Naïve Bayes perform poorly with low accuracy, but it trained the data faster comparing to SVM.

KEYWORDS

NLP, Twitter, Chatbot, Machine Learning, Sentiment Analysis, SVM, Naïve Bayes

1. INTRODUCTION

As additional cutting-edge technology of the industry 4.0 has progressed further, the social media platforms are embedded into different applications. Social media and computer networks have a great importance for communication between people and understanding the public feelings [1]. Therefore, there is an essential need to have a chatbot that aids in processing and analyzing social media data to achieve the optimal use of these platforms.

The Covid-19 pandemic is one of the issues that preoccupied the world in the past couple of years, and had a significant impact physically, mentally, socially, economically [2]. Therefore, the social media interactions have increased with peoples' posts and comments that reflected their feelings and motivation towards the Covid-19 pandemic and its impact on their health and economic state. For example, during the pandemic, some people expressed their experiences and their panic when they got sick while others expressed their opinions toward having the vaccinations. Moreover, many politicians and decision makers from different positions shared with their perspectives toward the procedures and precautions of this pandemic by using different social media platforms.

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 41-55, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130404

Multi studies and research have been conducted during the Covid-19 pandemic on peoples' posts through social media applications to understand people's feelings and interactions, and to know the most frequently questions along with phrases that were used during these times [3]. Sentiment analysis is an approach that creates a relation between different parts of text with sending emotions from those who post this particular text.

In this study, we will focus on developing a chatbot for sentiment analysis of the participants' tweets on twitter that relate to Covid-19 pandemic. We applied this study on a dataset of English tweets suitable for sentiment analysis, where the tweets would be classified as extremely positive, positive, neutral, negative, and extremely negative.

The purpose of this study is to have a chatbot which evaluates two algorithms of the machine learning that is used on sentiment analysis for participant's tweets related to Covid-19. We used the Support Vector Machines (SVM) and the Naïve Bayes (NB) algorithms to compare them based on the accuracy of the classifier and the execution time. Additionally, we then analyzed the difference of accuracy based by changing the test split ratios in both classifiers. Figure 1 shows the diagram of processing the sentiment analysis.

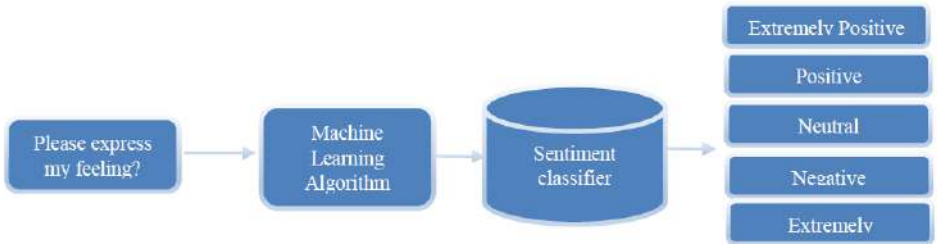


Figure1. The Framework of the chatbot of the sentiment analysis.

2. RELATED WORK

Researchers around the world are inspired to develop the state of-the-art chatbots by embedding different machine learning algorithms such as naïve Bayes algorithm and support vector machine (SVM) [4, 5].

Sentiment analysis is based on what people analyze, feel, and think [6]. Moreover, some authors perform mathematical calculations to examine people's feelings about a particular event and destination [7].

Rani & Singh [8]conducted a sentiment analysis for Twitter data which was collected by Twitter Application Programming Interface (API). Once they completed preprocessing the data, they used SVM for sentiment analysis with applying the following features: TF-IDF, Linear, and Kernel. However, they used F-score, recall, accuracy, and precision in order to measure the performance. The results revealed that linear SVM was given higher accuracy than Kernel SVM. Alabid, &Katheeth[9]used SVM to predict the sentiment analysis for twitter data that are related to social distancing during the COVID-19 pandemic. They used recall, F1, precision, and confusion matrix in order to evaluate the performance of the SVM algorithm.

They applied in their study 629 tweet texts and divided it as the following: 40% of tweets showed neutral sentiments, 25% of tweets showed positive, while 35% of tweets showed negative. This was followed then by dividing the dataset to 80% training and 20% testing data. After applying the SVM algorithm, the result of the performance evaluation of accuracy was 71%, but when it

was applied on the positive and negative tweet texts only, the percentage of accuracy was increased to 81%. As well as when they reduced the test data to 10%, they observed that the accuracy increased to 87%.

Finally, it was shown that increasing the training data would increase the performance of the algorithm [9].

Naw [10] used SVM and K-Nearest Neighbor (KNN) algorithms to conduct sentiment analysis on dataset collected by Twitter API. The author used the term frequency - inverse document frequency (TF-IDF) as a feature selection for classification, and after applying SVM and KNN algorithms, the data were classified as negative, neutral, and positive [10].

Alabid & Katheeth [9] used SVM and Naïve Bayes algorithms to conduct a sentiment analysis of the twitter texts related to the COVID-19 vaccines. The ratio of training data was 80% and the ratio of testing data was 20%. They preprocessed the dataset by removed stop words, punctuation, and tokenization. In addition, they applied part -of-speech (pos) tag. Subsequently, they selected the adjectives sentences which help to clear the ambiguous words. Through the results, it was found that SVM was better than NB with test ratio .01 while the stop words was removed from the texts. On the other hand, the results showed that the performance of NB was better than SVM with ratio .06, when they used PoS tag in addition of removing stop words. Also, other preprocessing techniques were applied as well to process unstructured twitter texts [9]. In general, sentiment analysis attracted a lot of researchers to pay more attention to this field and to use several algorithms to improve the classifiers.

Indeed, social media played a significant role during the Covid-19, driving researchers around the world to use several techniques in Natural Language Processing (NLP) to analyze people's perspectives and experiences during this pandemic.

2.1. Sentiment Analysis Based on Social Media Posts During Covid-19

Ouerhani et al. [11] developed a chatbot, called COVID-Chatbot to communicate with people during Covid-19 to increase their consciousness towards the real danger of this pandemic.
Liu et al. [6] conducted a research paper to study how people think and behave during the Covid-19 pandemic from the lens of social media posts by using the BERT (Bidirectional Encoder Representations from Transformers), as well as the clustering techniques.

In general, most studies that were conducted were intended to study the people's feeling in order to measure and detect their anxieties and depressions. Fauziah et al. [12] developed two machine language algorithms, the random forest and xgboost in order to detect the anxiety feeling during the pandemic, where the author used 4862 records from a dataset that was collected from YouTube comments. Moreover, [13] used the Machine Learning for detecting the patients' anxiety during the Covid-19 pandemic by using data from two different types of social media apps namely a communication app as well as a social networking app. On the other hand, [14] used Facebook's dataset in order to predict the spreading of new cases of Covid-19.

Chin et al. [15] analyzed 19,782 conversation utterances that are related to COVID-19 which cover different countries between 2020 and 2021. The authors identified chat topics (NLP) methods to analyze the emotional sentiments.

Several researchers conducted a sentiment analysis particularly during Covid-19 pandemic by using tweets datasets and different machine language models. Yao et al. [16] and other authors used advanced machine language algorithms to detect peoples' interaction from the vaccinations

[17] [18]. Support Vector Machine has been used from different authors to measure the sentiment analysis, for example Hayati et al. [19] and Sabrila et al. [20] used the Support Vector Machine algorithm as well as K-Nearest Neighbor Algorithm.

Table 1 shows several works have been done during the Covid-19 pandemic to predict peoples' interactions and behaviors by using different Machine Learning algorithms and Natural Language Processing.

Table 1. Several studies during Covid 19 for predicting people's interactions by using NLP & ML

Author	**Social Media**	**Sentiment Analysis Approach**
[15] (Chin et al. 2022)	SimSimi, one of the world's largest open-domain social chatbots	Natural language processing (NLP) methods
[11] (Ouerhani et al. 2020)	Utterances/ Ongoing Discussion	Natural language processing (NLP)/ Deep Learning/ LSTM
[6] (Liu et al. , 2021)	Reddit posts	BERT-based (Bidirectional Encoder Representations from Transformers)
[12] (Fauziah et al. ,2020)	YouTube	Random forest and xgboost
[21] (Li et al., 2022)	Sina Weibo, a leading social media platform in China.	NLP techniques and Regression Analysis
[13] (Ryu et al. ,2021)	Social media apps (communication and social networking)	Markov model and logistic regression
[22] (Tekumalla and Banda, 2020)	Twitter	NLP and ML
[23] (Sivanantham, 2021)	Web Comments and Blogging	SVM, logistic regression, and neural network
[24] (Bernado et al. ,2021)	Twitter	Naïve Bayes and Support Vector Machine
[25] (Ali, Malik,andMaheen 2021)	Twitter	Naïve Bayes, Logistic Regression, SVM, Deep LSTM, and BERT
[26] (Kumaresh, 2021)	Twitter	Naïve Bayes and Logistic regression

3. METHODOLOGY

This study adopted a scientific approach by using different independent and dependent variables in order to build a sentiment classifier. However, the methodology will be divided into four sections, 1- Dataset Selecting, 2- Data Preprocessing, 3- Training the Data, and 4- Testing the Machine Learning Algorithms. The Machine Learning (ML) model is developed by using Python software to import the ML packages such as the Scikit-learn, due to the fact that it's considered as one of the most powerful text processing tools that support and provide tokenization, filtration of tokens, and stemming.

3.1. Dataset Selecting

The Corpus is always our starting point for any Text-Pre-processing function, since it's the domain for our work, as it has the documents and documents have paragraphs, paragraphs include sentences, and each sentence is divided into words or what we can call, a token.

This study used a Tweets corpus as a collection of text tweets that were collected from the Twitter platform. In fact, the dataset that was used in this study is collected during the pandemic of Covid-19 and it's retrieved from Kaggle's website in CSV, and it includes 41157 tweets, and all are labeled and classified based on the sentiment of the tweet (Extremely Positive, Positive, Neutral, Negative, Extremely Negative). Moreover, the testing data split it into different ratios 10%, 20%, and 30% in order to compare the performance of SVM and Naïve Bayes models. Figure 2 shows the description for the tweets dataset.

```
[3]: train_dataset.describe()
```

[3]:

	UserName	ScreenName
count	41157.000000	41157.000000
mean	24377.000000	69329.000000
std	11881.146851	11881.146851
min	3799.000000	48751.000000
25%	14088.000000	59040.000000
50%	24377.000000	69329.000000
75%	34666.000000	79618.000000
max	44955.000000	89907.000000

Figure.2 Description for the dataset.

3.2. Data Preprocessing

Text or Data Pre-processing is an essential step for any Natural Language Processing system (NLP) since most elements of the texts such as characters, words, and sentences are important through the entire stages of the text processing. The purpose of all these stages is to make the text more analyzable for any particular task. Thus, in simple words, we can define it as a technique for converting the raw data into an understandable text, having only the meaningful words, that can be used for training the machines effectively.

In general, preprocessing the text includes four main processes: (1) Text Tokenization, (2) Removing stop words, (3) Normalization, and (4) Stemming/ Lemmatization. These four processes are utilized in order to simplify the text to a new format that can be utilized by NLP applications.
NLTK (Natural Language Toolkit) in Python, is the most important component for preprocessing the text. We used this library for almost all preprocessing functions.

3.2.1. Text Tokenization

The Tokenization -or some researchers call it as a segmentation –is the first step for NLP preprocessing the text and it's defined as splitting the text into characters, words, symbols, or sub-words (combination of words) as a token by using different techniques. However, the sub-words known as n-grams and (n), are considered as number of tokens, since some words can be more understandable when combined. In fact, the Tokenization has a significant impact on analyzing and processing any texts in terms that these tokens become as an input to other functions such as parsing and data mining. Moreover, an effective tokenization can play an essential role in reducing the input text documents and other actions that would be involved in NLP processing.

It was found that the tokenization technique is very effective in NLTK, we used both the word tokenizer and sentence tokenizer for tokenizing the datasets.

3.2.2. Removing Stop words

The datasets for both the Training and the Testing are cleaned from the Stop words by importing the module by using this code "from NLTK corpus import stop words" from NLTK library, to maximize the efficiency of the Dataset.

3.2.3. Normalizing and Stemming/ Lemmatization the Data

Stemming and Lemmatization could be related to Normalization in terms of simplifying the words to a unique meaningful word, since one word can turn into different forms of the word, but all can be shared by the same meaning. For example, "work", "works", "working", "worked", etc. without stemming and lemmatization the corpus will be tokenize as 4 different tokens, but after preprocessing it will be counted only one token" work".

3.3. Training the Machine Learning Model

In this study, we selected Python for deploying our two selected machine language algorithms, the Naïve Bayes algorithm and the Support Vector Machine, due to the fact that Python includes different Machine Learning libraries such as scikit-learn, TensorFlow, etc. Besides that, Python is the most preferred language for data science and machine learning due to the low-level libraries and clean high-level APIs. Moreover, we used the Jupyter Notebook 3.0.14 for coding and presenting the data, since it's considered as one of the most powerful environments for data scientists. In this study we used the "fit ( )" method from sklearn objects for fitting the model by using the training dataset, as the below code:

"pipeline. Fit (train dataset ['Original Tweet'], train dataset['Sentiment']) "

3.3.1. Naïve Bayes Machine Learning

Naïve Bayes algorithms was used for classification, a it is one of the supervised learning algorithms. This classifier works by training the data with the five below labeled categorical input: Extremely Positive 2- Positive 3- Neutral 4- Negative 5- Extremely Negative

This classifier works based on Bayes theorem by calculating the probabilities for each class. For example, in our dataset, we have positive and negative tweets. First, we need to classify whether each word in the tweet is Negative or Positive and then will calculate the frequency in each one. This would be followed by creating the probability for each class. Figure 2 shows a sample of positive and negative tweets, Table 2 explain how the Naïve Bayes algorithms work.

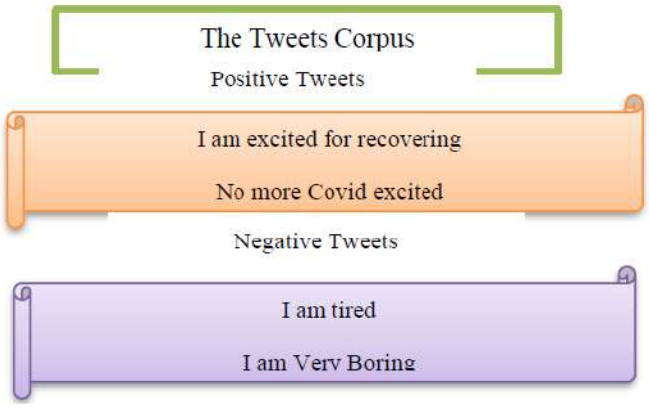


Figure2. Sample of Negative and Positive tweets.

Table 2. Shows the frequency table of Naïve Bayes algorithm

Frequency Table		
Words	**Pos**	**Neg**
I	1	2
am	1	2
excited	2	0
For	1	0
Recovering	1	0
Tired	0	1
No	1	0
More	1	0
Covid	1	0
Very	0	1
Boring	0	1
Total	9	7

3.3.1.1. Bayes Theorem

Naïve Bayes classifier is solved by using Bayes theorem:

P(A|B) = (P(B|A) * P(A)) / P(B)
P(A|B): The probability of event A given B (called posterior)
P(B|A): The probability of event B given A (called likelihood)
P(A): The probability of event A (called prior)
P(B): The probability of event B (called evidence)

We can apply it into the tweets predictions as the following:

P (Pos | “Recovering”) = P (“Recovering” | Pos) * P(Pos) / P (Recovering)

The word “Recovering” is a positive sentiment? Is this statement correct?
= (1/9 * 9/16) / (1/16)
= (.11 *.56) / .062 = .99

Naive Bayes uses a similar method to predict the probability of different class (Negative, Neutral, Extremely Negative, Extremely Positive).

In this study we used the below code as shown in Figure 3 for defining the Naive Bayes classifier and fitting the training dataset as the below code:

```
[83]: from sklearn.naive_bayes import MultinomialNB
      classifier = MultinomialNB()

[84]: from sklearn.pipeline import Pipeline

      pipeline = Pipeline([
          ('bow', bow),  # strings to token integer counts
          ('tfidf', tfidf),  # integer counts to weighted TF-IDF scores
          ('classifier', classifier),  # train on TF-IDF vectors w/ Naive Bayes classifier
      ])

[86]: pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[86]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x000001FC33DE6C10>)),
                      ('tfidf', TfidfTransformer()),
                      ('classifier', MultinomialNB())])
```

Figure 3. Fitting the training dataset by using SVM

3.3.2. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be applied in classification and regression analysis, however in this study we will use it in the classification model in order to predict the sentiment labels: Extremely Positive, Positive, Neutral, Negative, Extremely Negative. The idea behind SVM is finding a hyperplane that can best divide our training dataset into five different classes, which is known as (multiclass classification), however Figure 4 illustrated the five different classes.

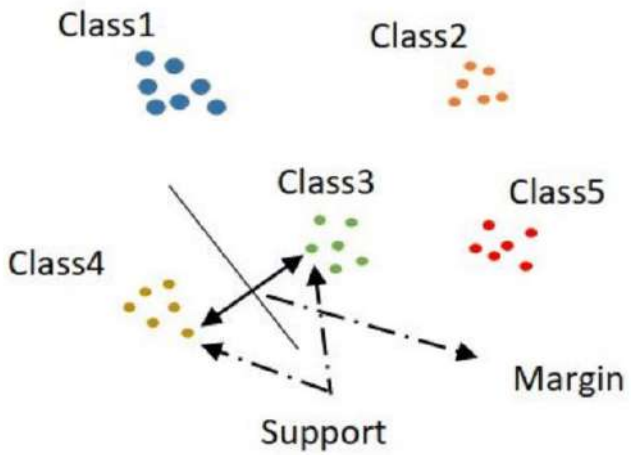


Figure 4. Shows the training technique by SVM

In this study we used the below code in Figure 5 for defining the SVM classifier and fitting the training dataset as the below code:

```
[23]:
from sklearn import svm
clf = svm.SVC()
classifier = clf

[24]: from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ('bow', bow),  # strings to token integer counts
    ('tfidf', tfidf),  # integer counts to weighted TF-IDF scores
    ('classifier', classifier),  # train on TF-IDF vectors w/ Naive Bayes classifier
])

[25]: pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[25]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x000001A929AC5700>)),
                      ('tfidf', TfidfTransformer()), ('classifier', SVC())])
```

Figure 5. Fitting the training dataset by using Naïve Bayes Classifier

3.4. Testing the Machine Learning Algorithms

There are four effective measures applied in this study, which all are based on confusion matrix output which are (True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN)). Machine learning prediction depends on the following formulas of the prediction scores:

Precision(P) = TP/(TP+FP)
Recall(R) = TP/(TP+FN)
Accuracy(A) = (TP+TN)/ (TP + TN + FP + FN)
F-Measure (Micro-averaging) = 2. (P.R)/(P+R)

In this study we used the "predict () "method from sklearn objects for predicting the target values from the testing dataset since this data is unseen and is not learned before. The below code is implemented into the Python:
"all predictions = pipeline. Predict (train dataset ['Original Tweet'])"

3.4.1. Results and Discussions

The results reveal high performance in the Support Vector Machine (SVM) model accuracy comparing to Naïve Bayes model, as it shown in Table 4 and Figure 6. Indeed, the accuracy factor is very vital in terms of evaluating the Machine Learning model and it can increase the credibility to any algorithm.

Therefore, in this study, we tried to improve the performance by changing the test split ratio as the following 10%, 20% and 30%. However, table 4 shows the results of the accuracy into the two algorithms.

Table 3. The accuracy results of SVM and Naïve Bayes models in changing the test split ratios

Machine Learning Model	**Test Split Ratio Accuracy**		
	10%	20%	30%
Naïve Bayes	37%	36%	35%
SVM	97%	97%	97%

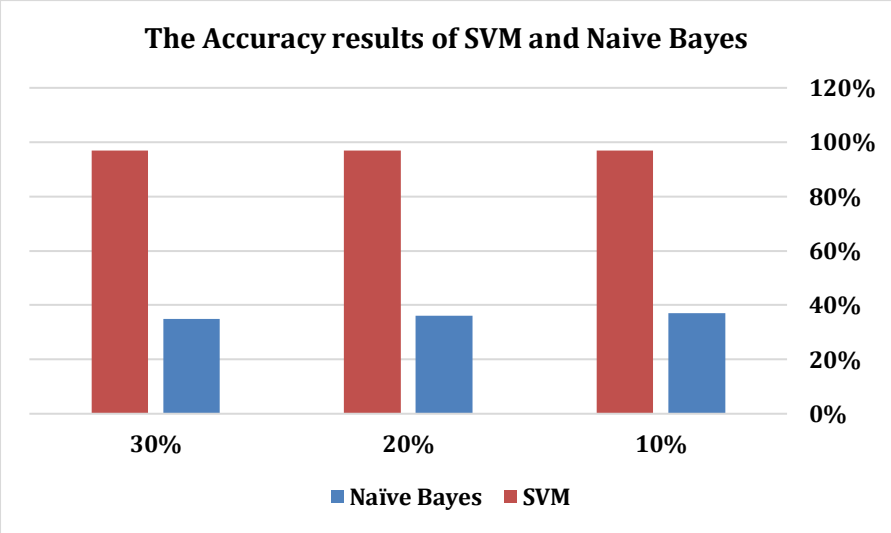


Figure 6. The Accuracy results of SVM and naïve Bayes

Moreover, as it shows in table 3, the models are more accurate when decreasing the test-split ratios in training the datasets. Appendix A shows the prediction codes along with the results by using different split ratios in SVM and Naïve Bayes models.

Table 4. The training speed per minutes in SVM and Naïve Bayes classifier

Machine Learning Model	**The speed per minutes / Split Ratios**		
	10%	20%	30%
Naïve Bayes	3 Minutes	2 Minutes	2 Minutes
SVM	40 Minutes	35 Minutes	35 Minutes

Moreover, the study reveals that training speed in the SVM classifier is relatively slow comparing to Naïve Bayes classifier. Table 4 shows the training speed per minutes for both classifiers and by using different test split ratios. Figure 7 shows the chatbot predicting the feeling from the texts.

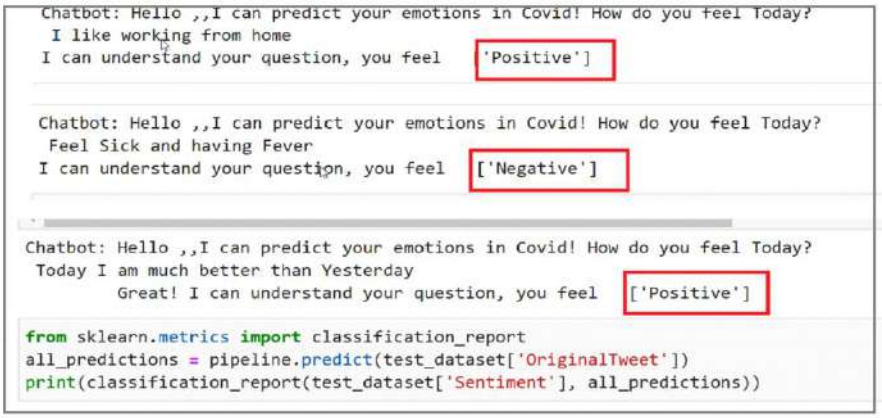


Figure7. Chatbot's prediction –predicting the feeling from the texts

4. CONCLUSIONS

The main goal of this paper is evaluating the accuracy performance of predicting the sentiment classes by training the tweets datasets in two models of machine learning algorithms SVM and Naïve Bayes and then evaluating the role of test split ratios in the accuracy performance. However, the results revealed that the accuracy increased when decreasing test split ratios. Also, the results showed a high performance in (SVM) model accuracy comparing with NB model. Moreover, the study revealed that training speed varied in both models, since the speed of SVM classifier is extremely slow even though it is more accurate classifier.

5. FUTURE STUDIES

In this study we worked only with SVM and Naïve Bayes algorithms. Therefore, the next step would be to explore to other algorithms such as the deep learning models. In addition to that, improving our algorithms by studying the role of features that could have positive impact on the speed of the SVM classifier without affecting the accuracy of the predictions is something we would like to focus on in the future. And lastly, this study only focused on the English tweets and because of that, for the future, we will improve it by including other languages such as the Arabic language.

REFERENCES

[1] Kim E.H.-J., Jeong Y.K., Kim Y., Kang K.Y., Song M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science.* 2016;42:763–781

[2] Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta bio-medica: Atenei Parmensis.2020; 91(1):157–160

[3] Anstead, N. and O'Loughlin, B., 2015. Social media analysis and public opinion: The 2010 UK general election. *Journal of computer-mediated communication*, *20*(2), pp.204-220.

[4] Bird, J. J., Ekárt, A., &Faria, D. R. (2021). Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 1-16.

[5] Assayed, S. K., Shaalan, K., & Alkhatib, M. (2023). A Chatbot Intent Classifier for Supporting High School Students. *EAI Endorsed Transactions on Scalable Information Systems*, e1-e1.

[6] Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T. and Anwar, M., 2021. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, *9*(1), pp.1-16.

[7] Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. Ain Shams Eng. J. 2014, 5,1093–1113. [CrossRef]

[8] Rani, S., & Singh, J. (2017). Sentiment analysis of Tweets using support vector machine. *Int. J. Comput. Sci. Mob. Appl*, *5*(10), 83-91.

[9] Alabid, N. N., &Katheeth, Z. D. (2021). Sentiment analysis of Twitter posts related to the COVID-19 vaccines. *Indonesian Journal of Electrical Engineering and Computer Science*, *24*(3), 1727-1734.

[10] Naw, N. (2018). Twitter sentiment analysis using support vector machine and K-NN classifiers. *IJSRP*, *8*, 407-411.

[11] Ouerhani, N., Maalel, A., Ghézala, H. B., &Chouri, S. (2020). Smart ubiquitous chatbot for COVID-19 assistance with deep learning sentiment analysis model during and after quarantine

[12] Fauziah, Y., Saifullah, S. and Aribowo, A.S., 2020, October. Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic. In *Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series* (Vol. 1, No. 1, pp. 253-261).

[13] Ryu, J., Sükei, E., Norbury, A., Liu, S.H., Campaña-Montes, J.J., Baca-Garcia, E., Artés, A. and Perez-Rodriguez, M.M., 2021. Shift in Social Media App Usage During COVID-19 Lockdown and Clinical Anxiety Symptoms: Machine Learning–Based Ecological Momentary Assessment Study. *JMIR mental health*, *8*(9), p.e30833.

[14] Vahedi, B., Karimzadeh, M. and Zoraghein, H., 2021. Predicting county-level COVID-19 cases using spatiotemporal machine learning: Modeling human interactions using social media and cell-phone data.

[15] Chin, H., Lima, G., Shin, M., Zhunis, A., Cha, C., Choi, J., & Cha, M. (2022). User-Chatbot Conversations During the COVID-19 Pandemic: A Study Based on Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*.

[16] Yao, Z., Yang, J., Liu, J., Keith, M., & Guan, C. (2021). Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19. *Cities*, *116*, 103273.

[17] Kwok, S.W.H., Vadde, S.K. and Wang, G., 2021. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. *Journal of medical Internet research*, *23*(5), p.e26953.

[18] Wibowo, D.A. and Musdholifah, A., 2021, December. Sentiments Analysis of Indonesian Tweet About Covid-19 Vaccine Using Support Vector Machine and Fasttext Embedding. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 184-188). IEEE.

[19] Hayati, H., &Alifi, M. R. (2021). Analis is sentiment pada tweet terkait vaksin Covid-19 menggunakan metode support vector machine. *JTT (Jurnal Teknologi Terapan)*, *7*(2), 110-119.

[20] Sabrila, T.S., Sari, V.R. and Minarno, A.E., 2021. Analisis Sentimen Pada Tweet Tentang Penanganan Covid-19 Menggunakan Word Embedding Pada Algoritma Support Vector Machine Dan K-Nearest Neighbor. *Fountain of Informatics Journal*, *6*(2), pp.69-75.

[21] Li, K., Zhou, C., Luo, X., Benitez, J. and Liao, Q., 2022. Impact of information timeliness and richness on public engagement on social media during COVID-19 pandemic: An empirical investigation based on NLP and machine learning. *Decision Support Systems*, p.113752.

[22] Tekumalla, R. and Banda, J.M., 2020. Characterization of potential drug treatments for covid-19 using social media data and machine learning.

[23] Sivanantham, K., 2021. Sentiment Analysis on Social Media for Emotional Prediction During COVID-19 Pandemic Using Efficient Machine Learning Approach. *Computational Intelligence and Healthcare Informatics*, pp.215-233.

[24] Bernado, J., Johnston, H., Powers, M., Xie, Y., Wang, J. and Gagnon-Bartsch, J., Public Opinions on COVID-19 Vaccines from Social Media: A Machine Learning Study.

[25] Ali, G., Malik, K.I. and Maheen, U., Comparative Analysis of Machine Learning and Deep Learning Algorithms for Classification of Social Media data related to COVID-19.

[26] Kumaresh, S., Sentiment Analysis Of Covid-19 Vaccine In A Social Media Platform Using Machine Learning Techniques. *Syndicate-The Journal Of Management*, P.39.

Appendix A

1- Naïve Bayes model

a -Split ratio to 10%

```
pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])
```

```
Pipeline(steps=[('bow',
                 CountVectorizer(analyzer=<function preprocess at 0x000001ED4DB214C
                ('tfidf', TfidfTransformer()),
                ('classifier', MultinomialNB())])
```

```
from sklearn.metrics import classification_report
all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
print(classification_report(test_dataset['Sentiment'], all_predictions))
```

```
                    precision    recall  f1-score   support

Extremely Negative       1.00      0.01      0.03       574
Extremely Positive       0.86      0.03      0.05       653
          Negative       0.40      0.43      0.41       991
           Neutral       0.80      0.05      0.10       756
          Positive       0.35      0.91      0.50      1142

          accuracy                           0.37      4116
         macro avg       0.68      0.29      0.22      4116
      weighted avg       0.62      0.37      0.27      4116
```

b- Split ratio to 20%

```
[60]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x000001ED4F03BD30>)),
                      ('tfidf', TfidfTransformer()),
                      ('classifier', MultinomialNB())])
```

```
[62]: from sklearn.metrics import classification_report
      all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
      print(classification_report(test_dataset['Sentiment'], all_predictions))
```

```
                    precision    recall  f1-score   support

Extremely Negative       0.96      0.01      0.03      1672
Extremely Positive       0.91      0.02      0.04      2018
          Negative       0.40      0.39      0.40      2985
           Neutral       0.81      0.05      0.09      2310
          Positive       0.33      0.90      0.48      3363

          accuracy                           0.36     12348
         macro avg       0.68      0.28      0.21     12348
      weighted avg       0.62      0.36      0.25     12348
```

c- Split ratio to 30%

```
[131]: Pipeline(steps=[('bow',
                        CountVectorizer(analyzer=<function preprocess at 0x000001ED59342160>)),
                       ('tfidf', TfidfTransformer()),
                       ('classifier', MultinomialNB())])

[132]: from sklearn.metrics import classification_report
       all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
       print(classification_report(test_dataset['Sentiment'], all_predictions))
```

	precision	recall	f1-score	support
Extremely Negative	0.86	0.01	0.01	3297
Extremely Positive	0.87	0.02	0.04	3959
Negative	0.39	0.38	0.38	5917
Neutral	0.83	0.03	0.07	4644
Positive	0.33	0.90	0.49	6878
accuracy			0.35	24695
macro avg	0.65	0.27	0.20	24695
weighted avg	0.59	0.35	0.25	24695

Fig 4. Prediction scores after using the Naïve Bayes algorithm

2- Support Vector Machine (SVM)

a- Split ratio to 10%

```
[25]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x000001A92
      9AC5700>)),
                      ('tfidf', TfidfTransformer()), ('classifier', SVC())])

[26]: from sklearn.metrics import classification_report

      all_predictions = pipeline.predict(train_dataset['OriginalTweet'])
      print(classification_report(train_dataset['Sentiment'], all_predictions))
```

	precision	recall	f1-score	support
Extremely Negative	0.98	0.96	0.97	5481
Extremely Positive	0.98	0.96	0.97	6624
Negative	0.95	0.98	0.97	9917
Neutral	0.99	0.95	0.97	7713
Positive	0.95	0.99	0.97	11422
accuracy			0.97	41157
macro avg	0.97	0.96	0.97	41157
weighted avg	0.97	0.97	0.97	41157

b- Split ratio to 20%

```
[46]: pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[46]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x0000019675D05940>)),
                      ('tfidf', TfidfTransformer()), ('classifier', SVC())])

[49]: from sklearn.metrics import classification_report

      all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
      print(classification_report(test_dataset['Sentiment'], all_predictions))
```

	precision	recall	f1-score	support
Extremely Negative	0.98	0.95	0.96	1142
Extremely Positive	0.99	0.95	0.97	1314
Negative	0.95	0.98	0.97	1939
Neutral	0.99	0.95	0.97	1510
Positive	0.95	0.99	0.97	2327
accuracy			0.97	8232
macro avg	0.97	0.96	0.97	8232
weighted avg	0.97	0.97	0.97	8232

Split ratio to 30%

```
[75]: pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[75]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x0000019675B44310>)),
                      ('tfidf', TfidfTransformer()), ('classifier', SVC())])

[77]: from sklearn.metrics import classification_report

      all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
      print(classification_report(test_dataset['Sentiment'], all_predictions))
```

	precision	recall	f1-score	support
Extremely Negative	0.98	0.96	0.97	1623
Extremely Positive	0.98	0.96	0.97	1981
Negative	0.95	0.97	0.96	2934
Neutral	0.99	0.95	0.97	2309
Positive	0.95	0.98	0.97	3501
accuracy			0.97	12348
macro avg	0.97	0.96	0.97	12348
weighted avg	0.97	0.97	0.97	12348

AN INTEGRATIVE APP PRODUCING AN OPTIMAL PATH FOR THE VESSEL IN ORDER TO REDUCE THE IMPACTS OF CARGO SHIPS ON THE ENVIRONMENT

Chenyu Zuo[1], Yu Sun[2]

[1]Sage Hill School, 20402 Newport Coast Dr, Newport Beach, CA 92657
[2]California State Polytechnic University, Pomona, CA, 91768,
Irvine, CA 92620

ABSTRACT

Almost every business in the world relies in some way on the shipping industry, whether it is to ship goods or natural resources, the shipping industry is undeniably the global industry, However, these very ships that drive the economy also produce close to 1 billion metric tons of carbon dioxide per year. In this project, we explore the use of machine learning to improve the performance of cargo ships in the ocean by implementing a genetic algorithm AI and a virtual simulation environment. An app was made based on using the training developed by the AI to be able to be deployed on cargo ships as part of their navigation system. Once sufficient data regarding a vessel's environment was collected, the algorithm could then produce an optimal path for the vessel. Experiments show that the AI system could sufficiently adjust to varying conditions and produce optimal paths for vessels.

KEYWORDS

Machine Learning, AI, Mobile APP, environment

1. INTRODUCTION

With the expanding global supply chain that supports the world economy, cargo ships have become an important aspect of everyday life. Almost 80% of world trade is handled by cargo ships which are equal to around 70% of the world trade value each year [1]. Goods from clothing to cars, from food to oil are all transported by sea, holding up the global economy and supporting the lives we live every day. Moreover, the global shipping value in 2019 reached 14 trillion US Dollars [2]. Despite the importance and growth of this industry, its environmental impact on cargo ships is often ignored. Cargo ships released 1,000 Mt CO2 per year, 3% of global CO2 emissions in 2020, and the rate of pollution is set to increase by 120%. While most of this pollution is created through the transport of goods, another area of concern is the pollution caused by cargo ships when they sit idle in traffic jams or backups along trade routes and ports. An example of this phenomenon could be seen in The Ports of Long Beach and Los Angeles, which handled around 10.7 million TEUs in CY 2021 each, are producing 100 tons of smog, or particulate-forming nitrogen oxides per day [3][4]. This level of pollution equals to around more than 6 million cars and has led to an increased rate of asthma and an increased rate of fatality during the Covid-19 pandemic in regions near or next to the ports [5]. This leaves our society in a dilemma, while the shipping industry is tremendously important to both the global economy and daily life, but also is actively destroying the environment and the health of everyday people.

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 57-69, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130405

Existing methods of reducing the impacts of cargo ships on the environment exist in two forms [6]. One, certain shipping companies have started to adopt a strategy referred to as "slow steaming", where ships are purposely sailed at a lower speed than the max speed to save money on fuel, which could slash fuel conception by 59%. [7]. However, this method is not very economically sustainable for most companies, as slower ships mean longer trips, which causes crew wages, maintenance costs, and other expenses to rise [8]. Meaning that slow steaming sometimes becomes financially unsustainable for shipping companies, who are already losing money due to the Covid-19 pandemic. Another solution was to create alternative-powered boats, such as boats relying on wind power or green fuel. Although these technologies do exist, the problem is that investing in these technologies is not commercially viable for many companies, meaning most of these technologies are not mature yet for commercial use. Another challenge is making sure that these solutions don't inherently cause more pollution, as in the case of alternative sources of fuel, which could produce less immersion initially, but produce more dangerous pollutants or be riskier to use on a large scale [9]. For example, some alternative sources of fuel suggested are hydrogen, ammonia, or biofuel, however, each of these fuels has its issues regarding safety such as toxicity or increased risk of fires and price, and are just generally more expensive to use for shipping companies. Moreover, even if these innovations do catch on, it would take years for new ships to be built and adapted by most of the world. In the meantime, a solution is needed that is both low risk both financially and in real life, well also effective enough to significantly reduce pollution caused by cargo ships.

In this paper, we will follow the line of research of "slow steaming" by making ships use less fuel by being efficient. However, instead of going slower, we are exploring if ships could be more efficient by using an Artificial Intelligence model to be more efficient in the water to go faster. By taking existing data that ships have on board regarding the conditions of the weather the AI system is applied to the current conditions to generate the most efficient pass for the boat to take. Inspired by the automated cars developed by some of the biggest companies in the world, the use of artificial intelligence to direct a transportation system is not something new. But, these systems are not designed to be the most efficient, nor are they on the water. First, sensors currently already installed in the modern cargo ship gather data regarding conditions which allows for the system to generate a simulation model reflecting conditions the boat is facing in real-time. Then, the reinforced learning system is applied to the simulation to find the best path for the boat. Then, the system sends the path generated to the control system, where the Captain could apply the path manually through a guidance system, or allow the autopilot to apply the path. Through this, we believe that fuel use could be cut drastically through a financially convenient and easy-to-install system that just needs a computer.

To prove our results, we will demonstrate the effectiveness of the above combination of techniques by comparing it to a regular path used by cargo ships and afterward, compare the distance and average speed to calculate the amount of fuel used during the trip to measure the fuel savings. First, we would find a set route for the boats to travel. Then, by first letting the simulation run the route using our simulation, and without it, we can gather data on the average speed and time needed for travel. Afterward, using the distance and average speed, the amount of fuel used could be gathered by using the average fuel used by cargo ships. Finally, the comparison of the fuel use would allow the demonstration of the effectiveness of the system.

This paper is organized as follows: Section 2 gives details and a list of challenges that we faced during the experiment, design, and creation of the product; Section 3 focuses on the details of our solutions corresponding to challenges that we mentioned in Section 2; Section 4 presents the relevant details and regarding the experiment along with data gathered and conclusions we drew through the data. Following section 4, section 5 presents the related work in Section 5. Finally,

Section 6 gives the conclusion remarks regarding the project and the future work and direction of this project.

2. CHALLENGES

To build the project, a few challenges have been identified as follows.

2.1. How to Build a Realistic Boat Simulation Environment

The first step to building an AI simulation is to build the simulation itself [13]. In our case, that means constructing an accurate boat model and an accurate model that could represent the ocean using data that average ship sensors could gather. However, figuring out how to utilize the tools available in Unity to represent certain conditions and variables in the water is the most important part. The issue is the number of conditions to take into account makes influencing the simulation immensely complicated. The main conditions we took into account are wind conditions, currents, wave conditions, and length of the path. To overcome this, we have combined the effects of the conditions into one single angular tilt of the simulation, which allows us to simulate accurately.

2.2. How to Integrate AI Algorithms in the Simulation Environment

The second challenge encountered when developing the simulation is integrating the AI algorithms into the simulation environment. In the simulation environment, there are two sets of coordinates, the global coordinate and the local coordinate. The global coordinate is the coordinate system of the whole environment, while local coordinates represent the individual simulations. For the simulations to function, the coordinate systems have to match with the AI algorithm, and by adjusting the position of the individual simulations, we connect the AI algorithm with the simulation environment, which allows for the program to function.

2.3. How to Improve the AI Algorithms without Depending on the Specific Environment Settings

Once the AI algorithm is integrated, however, more challenges arise, as through testing and data gathered, there seems to be no significant improvement in terms of paths the AI is generating. After investigating, it seems that the equation used to generate the punishment of the path was not sufficient to improve the performance of the path in a significant and timely manner. The issue with the old equation is that it only takes into account the vessel's final x position rewarding the target location, and not the x coordinate. So when we generate a new equation that does take into account both values, the system is able to better differentiate which path's and good and which are not, drastically improving performance.

3. SOLUTION

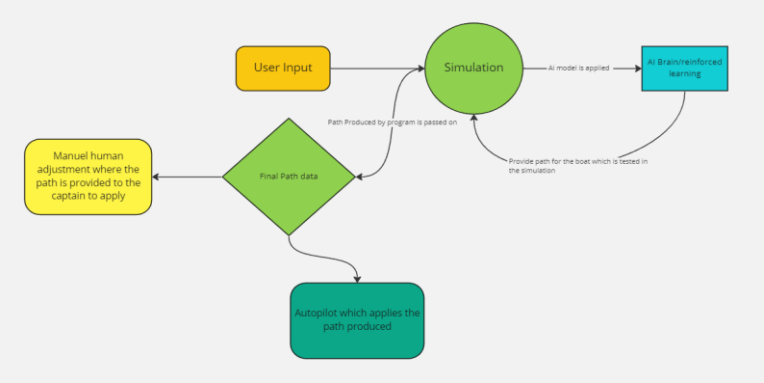


Figure 1. Overview of the solution

This diagram starts at the user input, where data such as wind speed, wave height and length, and current is entered into the system. Then, the data travels to the simulation, where they are used to adjust the simulation through different tilts and lengths of the simulation environment. Next, after the environment is finalized, the AI brain, (produced through reinforced learning), is applied to the simulation environment where it produces a path that is efficient and sufficient for the vessel. This path data is produced in the form of angles and speeds for the vessel to apply. Then there are two options, one where the path is automatically applied by the autopilot, which controls the speed and the direction of the vessel and could adjust the boat according to the new path directions, or the second option is a manual override, where the captain or senior officer of the ship could choose to apply the path manually or override it in favor of a different path.

The Components come together as a sort of cycle of information. Meaning that each passing and transformation of data directly affects the data produced in the next component. The entire system starts with the collection of real-world data in the form of weather and surface conditions. which is converted into a simulation environment built using Unity software. Then the AI brain is applied to the simulation which focuses on finding the optimal path for the particular environment. Afterward, that information is passed in the form of directions for autopilot or a captain.

The Unity Simulation environment in particular allows for the accurate simulation of conditions on the water, however, instead of directly simulating the conditions on the water which would have needed years of research, we are able to instead represent the effects of conditions in other ways such as tilts and speeds of the vessel. As seen in figure 2, the boat is represented as a white sphere that is launched at a constant speed across the platform. The platform is surrounded by white bars, which represent the boundaries and finish of the route. If the ball contacts the three walls behind it and to the left or right, it is counted as leaving the route meaning that it failed. The wall in front represents the finish or goal, which means that the path was able to reach the requested location efficiently. Using the application which we built, the system is able to take input, as shown in figure 3, in the form of boat speed (in knots), wave height and length, and travel distance, then translate the input into ball speed, platform tilt, and distance between the goal and ball respectively. This allows for both human input or a direct connection to an automated system that would collect data through sensors.

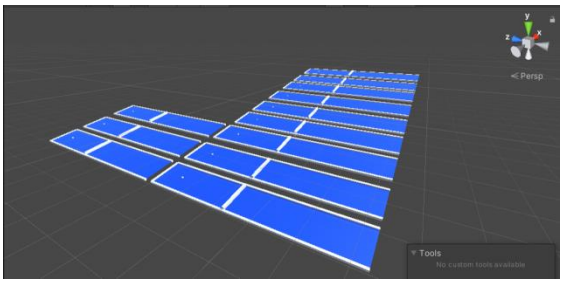


Figure 2. View of the simulation environments

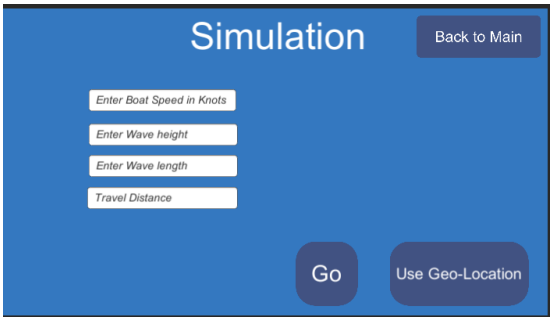


Figure 3. The input format for the application

When simulating, the ball would be launched through input produced by the Ml agents AI which would be further explained in the next section. This input comes in the form of two decimal numbers ranging from 0 to 1. This represents the angle at which the ball (vessel) would be launched, simulating the change in heading a vessel would undergo, then the speed would be adjusted according to the input. This can be seen in figure 4, first, the speed of the ball is set according to the vessel's speed, then the force is applied using input from the Genetic algorithm.

```
public override void OnActionReceived(ActionBuffers actions)
{
    //Debug.Log("OnActionReceived: " + " X: " + actions.ContinuousActions[0] + "  Y: " + actions.ContinuousActions[1]);
    // Debug.Log("X: " + actions.ContinuousActions[0]);
    // Debug.Log("Y: " + actions.ContinuousActions[1]);
    float moveX = actions.ContinuousActions[0];
    float moveZ = actions.ContinuousActions[1];

    // float moveX = 0f;
    // float moveZ = 1f;

    float moveSpeed = infos.b*100f;
    // Debug.Log("Foced: " + isForced);
    // Debug.Log(DateTime.Now - startTime);
    if (DateTime.Now > startTime.AddSeconds(10))
    {
        Debug.Log("Too slow! EndEpisode");
        SetReward(-1f);
        floorMeshRenderer.material = loseMaterial;
        EndEpisode();
    }
    if (!isForced)
    {
        Debug.Log("Applied the force!");
        Debug.Log("OnActionReceived: " + " X: " + actions.ContinuousActions[0] + "  Y: " + actions.ContinuousActions[1]);
        m_Rigidbody.AddForce(new Vector3(moveX, 0, moveZ) * moveSpeed);
        isForced = true;
    }
    // transform.localPosition += new Vector3(moveX, 0, moveZ) * Time.deltaTime * moveSpeed;
}
```

Figure 4. Code for action Simulator in accordance with input data

In order to construct the AI brain for the system, an AI simulation would be needed to be implemented first. A simulation allows the AI brain to "develop" in a way, or in more technical terms, gather data regarding paths for different types of simulated conditions which it can use to apply to paths it may or may not have seen before.

The most important part of this simulation would be the model of Artificial Learning chosen for this project. Due to the goal and needs of the simulation, a Reinforcement learning Algorithm was chosen to be created using the Ml agent AI system. This choice was made due to two factors. One, the Algorithm would smoothly integrate into the existing simulation model as described in section 3.2.1. Two, the system of the algorithm, which in simple terms is an award/punishment, means that we can quickly separate efficient and inefficient paths in the simulation. In order to interpret the algorithm, a couple of changes need to be made to the algorithm. First, the Ml Agents AI produces a path in the value of one x and on the y value using a t first random data. Then, that produced path is passed to the simulation which is able to two important pieces of data.

A key piece of this process is the ml agent Ai, which handles the actual learning aspect of the AI. In simple terms, ml-agents take environmental feedback in terms of awards and punishments to improve its future decisions. So at first, the agent might move randomly with no particular direction. However, as soon as the agent commits an action with leads to an award or punishment, it is able to shape its decision-making quickly. By going through multiple interactions of awards/punishments, the agent is able to quickly make decisions based on past “experiences”. Even if the environment changes, through past experience, the AI is able to still make decisions based on environmental data and improve with each new environment.

What makes this AI unique is its use of Reinforcement learning, which allows it to remember past actions and improve based on results. By assigning a positive value to desired behaviors and negative values to undesirable behaviors or outcomes, the algorithm is designed for seeking long terms and maximum awards. This allows the algorithm to focus on finding the optimal behavior quickly and works best in an environment where the agent is allowed to “explore” the environment on its own and create data regarding the environment. Which makes it perfect for our use. By making early mistakes that may lead the agent in the complete opposite direction as what the user wants, the vessel is able to quickly correct itself using feedback and find the optimal route.

In order to integrate into the AI algorithm, first the agent needs to be created, this agent would act as the vessel for our use and would carry out the testing and creation of the optimal path. In order to incorporate ml agents, there are several things we would first need to adjust. For one, we would need to decide what type of space type we want for our action, essentially, this decides whether the action produced is the whole number or a float (decimal). For this project, we choose a continuous space type, which allows getting a more diverse and accurate range of angles for the vessel to take. We also would adjust the max step to 15000, the max steps represent how many steps or paths in our case the AI is allowed to generate to find the optimal path. Then, by setting the actions to heuristic mode, we allow the AI to fully run on its own and begin testing.

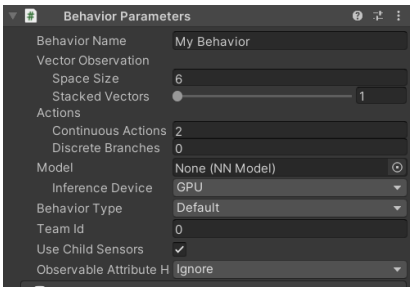


Figure 5. Behavior Parameters of the Agents

Once everything is set up, the agent still needs the reinforcement learning algorithm to learn from the simulation. This relies on the code seen in figure 6, which uses data collected from each

simulation, regarding both the position of the ball, and whether or not the path worked is adequate for the path specified from the start. Once the data is imputed, the code assigns an award or punishment based on the performance of the vessel, is the path is optimal and does reach the target location, it is assigned the award value of 1, however, if it does not reach the target location, then the punishment is calculated using the equation

$$P = -1 * ((z + L/2)/(L/2)) * ((x + w/2)/(w/2))$$

Where z and x represent the vessel's final position along the path z-axis and x-axis respectively, while L and W represent the goal's distance away from the initial position of the vessel and the width of the acceptable final location zone respectively.

The purpose of the equation is to further refine the process of training the AI to further develop better paths. Although all of the paths deemed unacceptable would not reach the target location, in the initial phases of training, there is a very low chance that the path generated would be successful, meaning that to speed up training, the AI would need to be able to improve upon the initially unsuccessful path. What the equation allows us to do is assign different levels of punishment to different levels of failure. For example, if the vessel ends up closer to the target than in previous attempts, then it would receive less punishment, essentially promoting that path above others. The initial separation of success would mean that the computer can more quickly create paths that have a higher chance to be successful.

```
private void OnTriggerEnter(Collider other)
{
    Debug.Log("Triggered!");
    if (other.TryGetComponent<Goal>(out Goal goal))
    {
        SetReward(+1f);
        floorMeshRenderer.material = winMaterial;
        Debug.Log("Win!");
        EndEpisode();
    }
    if (other.TryGetComponent<Wall>(out Wall wall))
    {
        float distance = (transform.localPosition.z + 31.3f) / 31.3f;
        float distance2 = ((transform.localPosition.x + 4) / 4f);
        distance2 = Math.Abs(distance2);
        // Debug.Log("Distance: " + distance);
        SetReward(-1f*distance*distance2);
        floorMeshRenderer.material = loseMaterial;
        Debug.Log("Lose!");
        EndEpisode();
    }
}
```

Figure 6. Reinforced Learning Algorithm

4. EXPERIMENT

In order to test the efficiency and accuracy of our system, we would need to ensure that the system works for all types of conditions on the water. This means simulating conditions commonly seen on the ocean, calm seas, and rough conditions. Furthermore, to test the flexibility of the system, we would also be testing the viability of the system to generate paths for longer distances. For those reasons, we would be testing the algorithm under three situations, a flat and short course, then we would be testing a longer course, and finally a longer course along with a larger degree of elevation. By running each condition through 15000 steps, we will gather data on the Accuracy and efficiency of the path.

4.1. Experiment 1:Flat and Short Course

For the first of our experiments, we used the standard design to simulate calm seas with a short distance of travel as seen in Figure 7. While not challenging for the human operator, the importance of the AI to be operated in any condition means even the calmest of easy need to be tested.

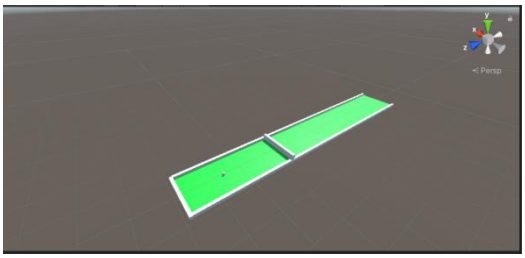


Figure 7. Flat and short course

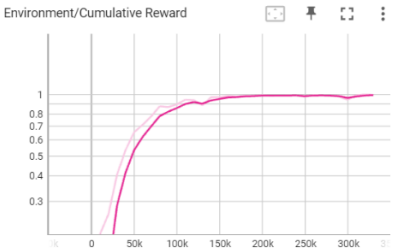


Figure 8. Environment/Cumulative reward

A. Accuracy

Despite an initial poor performance with an average reward of less than 0.3, the AI quickly finds an optimal path according to the judging system, reaching the upper limit of 1. This initial experiential growth demonstrates the AI quickly identifying which paths are optimal and which are not. Then, the AI was able to further pick out the most optimal path out of the chosen ones which can be seen through a curve from 50k to 100k step value. The resulting accuracy of close to 1 shows that the AI has successfully found the correct path in less than 150k steps.

B. Duration (steps)

A step, which is an atomic change of the engine that happens between Agent decisions, serves as our way of measuring the amount of work an agent or AI has spent developing an optimal path for a certain environment. Since the AI works very fast, steps instead of seconds are a better way for us to measure the performance of the agent.

C. Irregularities in the data

Certain irregularities can be found in the data for many reasons such as a slight decrease in performance. Such irregularities can be explained by the slow down of the computer or as the AI searches for more optimal paths, it can try to develop its own through existing paths, which can lead to improved performance or decreased performance as seen in these bumps.

4.2. Experiment 2

For the design of the second type of course, the width of the course and the position of the ball has not changed. However, instead of the finish being 31.5m away from the agent, the finish is moved to 51.5 m away shown in figure 9. This course is designed to test the flexibility of the program regarding distance and if it can adjust to the different needs of the captain, such as a longer target location.

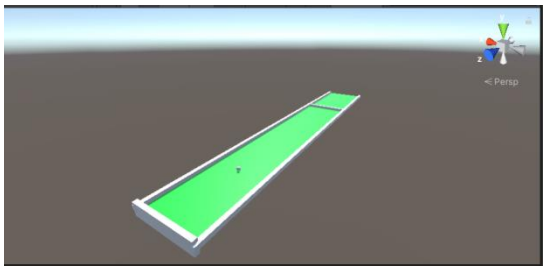


Figure 9. Flat and long course

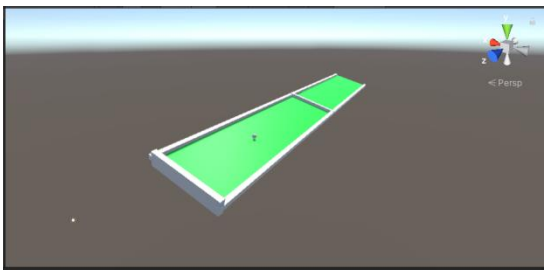


Figure 10. Tilted and Short Course

After the Simulation, the data provided in figure 10 showed another efficient performance. With a quick improvement, although slower than type 1, the AI is able to adapt to the different environment in roughly 150k steps, which as mentioned in section 4.1 are atomic changes made by the agent to attempt to adapt to the environment. The upper limit of 1 reached at 300k means that the AI has been able to find the most optimal path, which means that the AI is able to adapt to the environment adequately.

4.3. Experiment 3: Tilted Course

Type 3 courses represent rough conditions on the ocean which the AI would need to be able to adapt to in order to be adequate for service on vessels for use anywhere on the ocean. In order to simulate the tilt experienced cargo ships, the platform is tilted by 15 degrees as seen in figure 14.

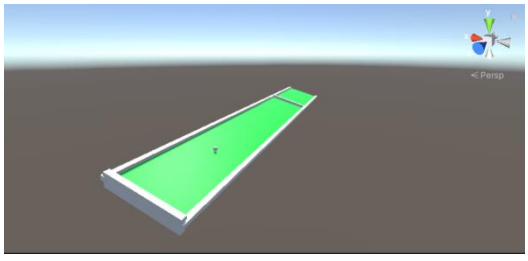


Figure 11. Tilted and long course

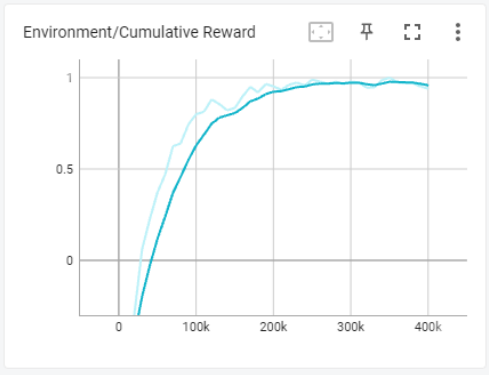


Figure 12. Environment/Cumulative Reward for type 3 simulation

After this simulation, the data in figure 11 shows that the AI was actually able to adjust to the tilted platform relatively quickly, finding optimal paths after 100k steps and reaching the upper limit of 1 around 225k steps. Again, the initial experiential growth represents the separation of non-optimal and optimal paths, while the curve before entering the upper limit represents the final improvement to the remaining paths. the Ai's ability to quickly adjust to different tilt would mean that it would be proficient in a real world setting.

4.4. Experiment 4

Type 4 being the last type is a combination of both type 2 and 3. This particular simulation serves as a stress test for the simulation. By combining both a further target distance and tilt, the environment tests if the simulation could efficiently adapt to both types of environments at the time time, and if the combination of conditions affect performances. In order to simulate the rough conditions and longer travel distances, the simulation environment has been tilted and extended to match the environments the experiment requires. So with a distance of 51.5 to cover and a 15 degree tilt as seen in figure 13, the environment would provide an adequate challenge for the ai

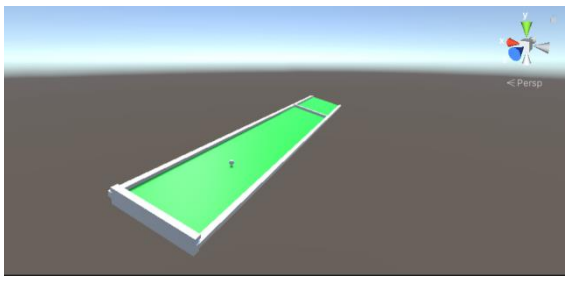


Figure 13. Tilted and long course

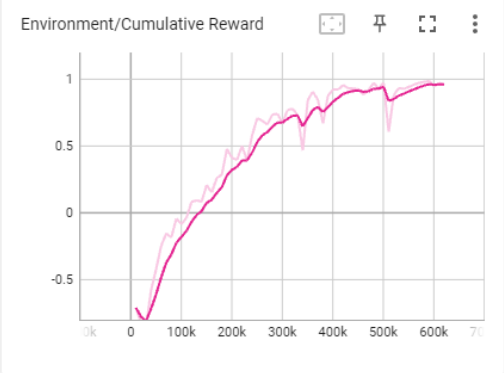


Figure 14: Environment/Cumulative Reward for type 1 simulation

As could be seen in Figure 14, the performance of the AI is certainly not as efficient as ones seen before, however, the AI is still able to find optimal paths after 250k steps and reach the upper limit of 1 around 450k. Which is performance expected for the complexity of the environment compared to previous ones. Although more irregularities in the data do exist due to the simulation's complexity, the ai's ability to reach the upper limit means that the AI is optimal and adequate for use on a real-world level.

Through the result shown through data collected on the four simulations, it becomes apparent that the machine learning algorithm can adequately respond to different types of environments. By finding the most optimal paths for a cargo ship, the ship is able to minimize its pollution impact but to use the same or even less fuel than before. Moreover, by being more efficient, the travel time gets cut dramatically over time, which further limits the pollution impact and lessens the strain on the global economy by reducing the chance of traffic jams due to late arrivals and allowing for better planning.

The design of the experiment with 4 different simulations not only ensures we cover all real world possibilities, but also ensures that we have enough samples to test the viability of the solution scientifically. By collecting trends in the graphs, we are able to notice the general trend of the growth for each environment. By measuring through their performance and steps, we are able to both evaluate general performance and irregularities that allow opportunities or improvement.

5. RELATED WORK

Zhang, M.et al presented a probability model that assesses the effects of human errors when utilizing an autonomous vessel [10]. They chose to focus on the relation between the human component of the human-autonomous interaction as part of an autonomous system. At the same time, our project focuses on the autonomous element. Moreover, while their paper focused on the probability of events that have already happened and analyzed performance, our paper focused on the creation of paths for a vessel to take which would be the most optimal. While the analysis of performance is essential when judging the performance, their work's focus on probability does not provide much insight into how the system could be improved but only highlights the problem through probability. Our project on the other hand focuses on finding a solution to the issue of inefficiency in relation to the environmental impact.

Kooj, C. et al investigated different effects regarding automated ships, and that is the effects on the crewing level due to the system [11]. Since an automated system would affect the tasks that the crew would need to handle, this paper investigates how an automated system would affect the

crew during sailing time, arrival, and departure of the cargo ships. This paper is an important aspect of developing new technology, especially regarding the shipping industry, as even though technology might be developing quickly through research like ours regarding the autonomous system itself, the effects of its implementation on ships and the shipping system as a whole can not be ignored.

Sui, S. et al investigate alternative methods of reducing the environmental impact and emissions of cargo [12]. The main methods they chose to focus on were alternative fuels such as liquefied natural gas (LNG) and the use of hybrid engines. While the focus of both of our papers is essentially on researching ways to reduce the impact of the shipping industry on the environment, the methods chosen are radically different. Their investigations focused on the idea of replacing emission-producing ships entirely with more environmentally friendly options. On the other hand, mine focuses on finding a way to improve current usable ships through an algorithm-driven AI which would make ships more efficient through their path.

6. CONCLUSIONS

In this project, we have proposed an algorithmic solution to the issue of environmental impact through cargo ships using machine learning and environmental simulation [14][15]. A platform designed to be used directly on shipboard computers has been proposed and created. Once a sufficient amount of data and training have been processed by the genetic algorithm, the system is able to take available data regarding conditions surrounding a vessel, and generate a path that would provide optimal efficiency, thus reducing fuel use, which leads to fewer emissions overall. Experiments show that the program can indeed adapt to different environments efficiently, and produce stable results quickly.

However, it is important that the technology produced in this paper has not yet been widely tested in real-world situations or adopted widely. Moreover, the technology does rely on a simulation that could be described as simple, which could be improved upon to be more accurate.

To address these issues, future work would be targeted toward improving the accuracy of the simulation and demonstrating the viability of our methods in real-world situations. In addition, improvements to the code and application to get them ready for real-world use would be needed. Features such as an interactive simulation or implementation of fluid dynamics for different models of ships would be needed.

REFERENCES

[1] Jacks, David S., and Krishna Pendakur. "Global trade and the maritime transport revolution." The Review of Economics and Statistics 92.4 (2010): 745-755.
[2] International Chamber of Shipping. "Shipping and World Trade: Driving Prosperity." (2021).
[3] Rosoff, Heather, and Detlof Von Winterfeldt. "A risk and economic analysis of dirty bomb attacks on the ports of Los Angeles and Long Beach." Risk Analysis: An International Journal 27.3 (2007): 533-546.
[4] National Research Council. Clean ships, clean ports, clean oceans: Controlling garbage and plastic wastes at sea. 1930.
[5] Mitchell, E. A. "International trends in hospital admission rates for asthma." Archives of disease in childhood 60.4 (1985): 376-378.
[6] Lipsitt, Jonah, et al. "Spatial analysis of COVID-19 and traffic-related air pollution in Los Angeles." Environment International 153 (2021): 106531.
[7] Wiesmann, Andreas. "Slow steaming–a viable long-term option." Wartsila Technical Journal 2 (2010): 49-55.

[8] Meyer, Jasper, Robert Stahlbock, and Stefan Voß. "Slow steaming in container shipping." 2012 45th Hawaii International Conference on System Sciences. IEEE, 2012.
[9] Kim, Hyungju, Kwi Yeon Koo, and Tae-Hwan Joung. "A study on the necessity of integrated evaluation of alternative marine fuels." Journal of International Maritime Safety, Environmental Affairs, and Shipping 4.2 (2020): 26-31.
[10] Zhang, Mingyang, et al. "A probabilistic model of human error assessment for autonomous cargo ships focusing on human–autonomy collaboration." Safety science 130 (2020): 104838.
[11] Kooij, Carmen, and Robert Hekkenberg. "Towards Unmanned Cargo-Ships: The Effects of Automating Navigational Tasks on Crewing Levels." Available at SSRN 3438144 (2019).
[12] Sui, Congbiao, et al. "Fuel consumption and emissions of ocean-going cargo ship with hybrid propulsion and different fuels over voyage." Journal of Marine Science and Engineering 8.8 (2020): 588.
[13] Fu, Daniel, and Ryan Houlette. "Putting AI in entertainment: An AI authoring tool for simulation and games." IEEE Intelligent Systems 17.4 (2002): 81-84.
[14] Ronen, David. "Cargo ships routing and scheduling: Survey of models and problems." European Journal of Operational Research 12.2 (1983): 119-126.
[15] Marans, Robert W., and Daniel Stokols, eds. "Environmental simulation: Research and policy issues." (2013).

AUTHORS

Chenyu Zuo is a second-year High school student at Sage Hill School located in Newport Beach, California. He is currently studying Multivariable Calculus and Physics. He is intreated in Coding, Sailing and wishes to one day explore the stars.

EYE-TRACKING IN ASSOCIATION WITH PHISHING CYBER ATTACKS: A COMPREHENSIVE LITERATURE REVIEW

Noon Hussein

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT

As of 2021, it has been reported that around 90% of data breaches occur on ac- count of phishing, while about 83% of organizations experienced phishing attacks [1]. Phishing can be defined as the cybercrime in which a target is contacted through e-mail, telephone or text message by someone impersonating a legitimate institution [2]. Through psychological manipulation, the threat actor attempts to deceive users into providing sensitive information, thereby causing financial and intellectual property losses, reputational damages, and operational activity disruption. In this light, this paper presents a comprehensive review of eye-tracking in association with phishing cyberattacks. To determine their impact on phishing detection accuracy, this work reviews 20 empirical studies which measure eye-tracking metrics with respect to different Areas of Interest (AOIs). The described experiments aim to produce simple cognitive user reactions, examine concentration, perception and trust in the system; all in which determine the level of susceptibility to deception and manipulation. Results suggest that longer gaze durations on AOIs, characterized by higher attention control, are strongly correlated with detection accuracy. Eye-tracking behavior also shows that technical background, domain knowledge, experience, training, and risk perception con- tribute to mitigating these attacks. Meanwhile, Time to First Fixation (TTFF), entry time and entry sequence data yielded inconclusive results regarding the impact on susceptibility to phishing attacks. The results aid in designing user-friendly URLs, visual browsing aids, and embedded and automated authentication systems. Most importantly, these findings can be used to establish user awareness through the development of training programs. be used to establish user awareness through the development of training programs.

KEYWORDS

Cybersecurity, Eye-Tracking, Phishing & Human Factors.

1. INTRODUCTION

According to recent security research, most companies have unprotected data and poor cybersecurity practices in place [3], which highly exposes them to security breaches. As the most common type of cyberattack, phishing describes the attempt to acquire sensitive information by disguising as a credible entity through e-mail, SMS, or phone. By creating a feeling of urgent necessity, inducing curiosity or fear in recipients, victims may reveal sensitive information, click on malicious links, or open attachments that may compromise their machines. As reported by the FBI's Internet Crime 3 (IC3), phishing was the most common cybercrime in 2020 [4]. In particular, one in every 99 e-mails is classified as a phishing e-mail, which makes it the most

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 71-85, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130406

common social engineering attack, comprising about 90% of security data breaches according to Cisco's 2021 Cybersecurity Threat Trends report [5].

Eye-tracking measures provide valuable non-invasive indices of human brain cognition. Based on gaze analysis, attentional focus and cognitive strategies are revealed. As the most commonly utilized ocular measure, eye gaze carries several advantages over EEG and fMRI for a number of paradigms and research questions.

Firstly, eye-tracking devices enable subjects to be comfortably seated or move freely with head-mounted devices during data collection. This results in a more natural and less space-restricted experimental environment compared to an MRI scanner. Secondly, since most eye-trackers are portable [6], it is easier to form larger and more diverse sample sizes, rather than being limited to subjects who are willing and able to commute to research facilities. Thirdly, the quick process of calibration on modern eye-trackers minimize pre-experiment set-up tasks and testing time.

Multiple gaze metrics used to assess cognition are derived from gaze position data. Gaze position measurements assess the thought process in a moment-by-moment manner for a variety of contexts. Fixations are used in the calculation of time spent looking at a particular location, which reflects engagement of attention as well as time required to process that stimulus. From these metrics, researchers can gain insights into memory [7], processes of mental computations and reading [8,9], in addition to problem-solving and learning strategies [10,11].

Modern web browsers embed tools to aid users in making informed security decisions. For instance, visual indicators can be found within URL bars, whereas SSL padlocks allow for judging the legitimacy of websites. Unfortunately, these indicators have only shown partial success at phishing prevention. Aside to that, poor usability may become advantageous to phishers when masquerading as legitimate sources. As earlier security indicators have proven ineffectiveness, they pose a higher risk of falling victim to phishing. This is compounded by the fact that most users consider security a secondary task [12], which affects the likelihood of noticing security indicators. Furthermore, some security indicators are only visible when the content is secure, which makes the absence of security indicators even less likely to be acknowledged.

Given the serious potential consequences of phishing cyberattacks, it has become of conspicuous interest to deepen one's understanding of the impact of exploited human cognition factors on these attacks, in order to minimize or mitigate their repercussions. In this light, this paper reviews the impact of eye-tracking, mainly including gaze position and associated metrics, on the susceptibility to phishing cyberattacks.

To the best of my knowledge, this is the first paper to review phishing susceptibility through the lens of eye-tracking. After thoroughly searching key academic databases, a full range of journal articles between 2012 and 2022 addressing the application of eye-tracking technology in phishing detection was systematically assessed. Based on rigorous selection criteria, 20 eligible articles were selected for final review, as this study develops a taxonomy built upon a comprehensive range of scholarly journals.

The remainder of this paper is organized as follows: methods for independent and dependent variable measurement are described in Section 2. Section 3 comprises the key findings of the literature, whereas discussions and implications are detailed in Sections 4

and 5, respectively. Limitations of the reviewed studies are presented in Section 6. Lastly, conclusions are reported in Section 7.

2. MEASUREMENT METHODS

Eye-tracking measures the point of gaze and eye motion relative to the head. An eye-tracker is therefore capable of producing a gaze path video and large quantities of physiological data related to attention as well as emotion. These devices come in a mobile or stationary format depending on the focus of the experiment. For example, glasses (mobile) can give insight on attention, response placement of products or other stimuli. To investigate the impact of factors extracted from eye-tracking on susceptibility to phishing, 20 empirical studies [13-32] were reviewed.

Phishing stimuli presented to users at random comprised the independent variable in the experiments, which were typically within-subject studies. Timestamp, gaze position relative to phishing stimuli (X and Y), position in eye-tracker field of view (X and Y), pupil size, and validity code of each eye are parameters measured by eye-trackers. From these, different measurement metrics were derived in the studies, whereas Areas of Interest (AOIs) were used to link them to parts of the used stimulus. For the reviewed work, Table 1 summarizes common AOIs used to evaluate susceptibility to phishing e-mails.

Table 1. AOIs for phishing e-mail detection based on eye-tracking.

AOI	**Description**
E-mail address	• Attacker disguises themselves as a trusted source. • Engagement is more likely with this form of deception, especially if the source is “familiar” to the user. • Domain or entire e-mail could be spoofed.
Subject line	Exploits urgency, personalization and pressure
Addressee	• Gathered background information about the victim can be used to personalize the attack, therefore increasing susceptibility. • May also be addressed through generalized information from a trusted organization in which they are inclined to comply
Instruction line	• Generally highly personalized to appeal to targets. • Persuasiveness is enhanced by source address spoofing and shortened URLs to hide the destination of the link. • Decisions are made based on previous experiences, biases, or beliefs.

Adding to the above AOIs, the National Cyber Security Centre described financial information, misspelling, threat, and urgency as elements identified in public guidance on possible phishing e-mail indicators [33]. In addition, more specific AOIs were established in the literature for phishing URLs, as described in Table 2.

Table 2. AOI for phishing URL detection based on eye-tracking.

URL AOI	Description
Scheme	• Captures the scheme component and corresponding delimiters. • HTTPS is mainly used as the scheme.
Authority	• Fully qualified domain name (e.g., www.google.com is the authority component of https://www.google.com). • Or has form user@host (e.g., www.google.com@evil.com). • Can be split into user and host AOIs corresponding to user and host components.
Rest	Captures the rest component.
Response	Captures participant response for phishing e-mail classification.
Visual	Captures visual targets other than the aforementioned AOIs, such as: • Trusted Digital Certificate indicator in the web page; lock icon with a green background. • SSL/TLS encryption indicator. • Content quality and information on page.

To measure AOIs, static and dynamic measurement metrics were used. Static metrics were studied in [13], [14] and [20-22], which include personal attributes, such as name, gender, age, income, experience, knowledge, biometric identities, and ethnicity. Although gender and age were somewhat considered, static metrics were not strictly taken into account in the literature, and may be considered as secondary metrics when measuring such physiological factor. Instead, Table 3 describes main dynamic metrics found in the literature.

Table 3. Measurement metrics for eye-tracking-based phishing detection.

Metric	Description
Time to First Fixation (TTFF)	Time taken to look at the first AOI.
Gaze position	Point of gaze; where one is looking.
Fixation count	Denotes interest in a particular content.
Number of regressions	• Number of times a participant returned their gaze to a particular spot, defined by an AOI. • Indicates that the area drew attention and needed further scrutiny.
Glance duration	• Denotes depth of processing. • Characterized by a threshold of 100 ms in [18] and 500 ms in [13].
Entry time and entry sequence	• Time and fixation number that an area was attended to, respectively. • Denotes ease of attentional capture.
Total dwell time	Time taken to fully analyze one item.
Total time	Total time taken to finish the experiment.
Questionnaire	• Related to personal static features, security knowledge and behavior, eye-tracking experience, and others. • One pre-task questionnaire in [18] assessed mood along six emotional states using a 10-point scale to neutralize it before the tasks. • A sample questionnaire can be found in the appendices of [21].

After acquiring gaze data, datasets were usually extended by considering additional metrics (described in Table 3) that build upon the fundamental data. Other metrics used in few studies include times of clicks [18], actions taken (e.g, deleted/archived mail or helpdesk notification) [29], memory [22,32], and pupil size [18], [29], [32] to evaluate user susceptibility from static

and dynamic metrics. It is to be noted that although static metrics were not necessarily primary metrics in these experiments, they contributed to unexpected, complex and inconsistent results in relation to susceptibility, as highlighted in the following sections.

Eye-tracking devices used in the experiments are classified into mobile and stationary devices [34]. As compared to mobile eye-trackers, stationary eye trackers can only be used in a laboratory. To further visualize the experimental setting, a brief comparison of used eye-trackers in terms of frequency and accuracy is presented in Table 4.

Table 4. Frequency and accuracy characteristics of eye-tracking devices in experiments.

Study	Device	Type	Frequency (Hz)	Accuracy (°)
[13]	iMotions SMI RED 500	Stationary	500	0.4
[14], [19]	Tobii Pro Glasses 2*	Mobile (glasses)	100 [14], unspecified [19]	0.62
[15]	Tobii Pro TX300*	Stationary	Unspecified	0.5
[16], [23]	Tobii Pro X2-30*	Mobile (screen	Unspecified [16], 30 [23]	0.4
[17]	Tobii Pro T60XL*	Stationary	60	0.5
[18]	Ergoneers Dikablis Glasses	Mobile (glasses)	60	0.3
[20]	Tobii 1750*	Stationary	100	0.5
[21]	iMotions The Eye Tribe Tracker*	Mobile (screen)	60	0.5
[27]	JINS MEME	Mobile (glasses)	Roughly over 100	Unspecified
[28]	EyeLink 1000 Plus	Stationary	60	0.5
[29]	EyeTech DS TM3	Stationary	60	0.5
[26], [30], [32]	Tobii T120*	Stationary	Unspecified [26], [30], 60 [32]	0.5

*Discontinued

From the table, it can be inferred that: (1) stationary and mobile eye-trackers are almost equally as popular for such experiment, with Tobii eye-trackers being the most used, and (2) eye movements were mostly recorded at 60 Hz, whereas (3) the majority of eye-trackers used had an accuracy of 0.5°.

3. KEY FINDINGS

The key findings of reviewed studies are summarized in Table 5, where some common themes were observed. Firstly, technical attributes, which are described as form and content-related aspects of crafted phishing attacks majorly impacted user behavior. Users paid most attention to salient design elements, spelling, URLs, sender's address, personalized content, interface and security indicators. As a result, their decisions were highly impacted by their perceived legitimacy of these attributes.

Secondly, personal attributes contributed to the correct identification of phishing attacks. As suggested by the literature, users of technical background, domain knowledge, experience, attention control and risk perception showed higher attentiveness levels, resulting in higher detection accuracy. In addition, contradictory to assumptions, agreeableness was found to have little to no impact on susceptibility to these attacks, as no distinguishable trends from eye-tracking results could conclude otherwise. Although personal attributes can be rather difficult to change, adequate training can decrease susceptibility to phishing

attacks. Specifically, training users to be more attentive and critical of phishing AOIs, even if short, enhances the ability to detect phishing cues.

Finally, although the initial fixation on an AOI differed depending on personal and technical attributes, findings suggest that glance duration was dominated by domain names in phishing e-mails and URLs in phishing websites. However, when facing warnings or threats, glance duration was found to be the highest among these. Experimental results contradict some assumptions that warnings and threats may divert attention away from important security indicators or pressure users into complying with attackers' demands. As supported by evidence, users have classified this type of information as less trustworthy, and were more attentive to cues in security warnings, which activated pattern matching mechanisms and induced a positive behavior towards phishing attacks. All in all, a user of a high detection accuracy is characterized by high attention control; spending more time looking at an AOI. Personal attributes have also resulted in secure behavior which contributed to phishing mitigation. Contrarily, TTFF, entry time and entry sequence data yielded inconclusive results regarding impact on susceptibility to phishing attacks.

Table 5. Key findings of reviewed studies.

Study	**Methodology and Sample Size**	**Key Findings**
[13]	Experiment, 22 participants	• Users spent less time looking at phishing indicators than expected. • Financial phishing e-mail indicators yielded the least frequent number of fixations and the least overall dwell time compared to those with misspelling, urgency, and threats. • Misspelling and threats were considered less trustworthy than financial and urgency indicators. • The presence of phishing indicators did not considerably affect the time spent looking at the rest of the e-mail. • The trustworthiness rating cannot be explained by the total time spent looking at phishing indicators.
[14]	Experiment, 25 participants (3 excluded)	• The best phishing e-mail could fool 40% of participants with a technical background. • Mainly, users looked at the body and header of an e-mail. • Knowledge and processing time are the two most important factors for identifying phishing e-mails.
[15]	Experiment, 23 participants	• The average detection error rate was 41.1%, in which 61 (30.5%) were false negatives, and 28 (21.4%) were false positives out of the 331 times the address bar was gazed. • Identification of phishing websites is improved by checking the address bar.
[16]	Experiment, 107 participants	• Experienced users attended and recognized more security-related information cues. • Situational information security awareness is not significantly impacted by agreeableness. • Instead, it is negatively influenced by contextual relevance and misplaced salience. • Salient design elements, such as logos and images divert attention from security cues more than plain text. • Users are more attentive to cues in security warnings, which activate pattern matching mechanisms. • Perceiving phishing as threatening generates a fear that indirectly but strongly invokes taking protective actions.
[17]	Case study , 160	• Users paid more attention to AOIs rather than uninformative and

	participants	distraction areas. • According to the average pupil diameter, users paid least attention to the sender, and most to the main e-mail content, followed by the salutation. • Persistent highlighting reduced attention spans on the main content.
[18]	User study, 20 participants	• Users' cognitive resources have a cap of around 100 characters when vetting a URL. • Users tend to believe that the presence of "www" in the domain name indicates the safety of a URL, and do not carefully parse the URL beyond that.
[19]	User study, unspecified	• Users who are more susceptible to phishing mainly focus less informative components; the e-mail content and images (if present). • For those users, the total number of gazes is generally lower. • Users who are less susceptible to phishing focus more on the sender's address and the URL (if present).
[20]	User study, 21 participants	• When evaluating the authenticity of a website, users only spend 6% of their time looking at security indicators. • On the other hand, 85% of their time is spent looking at website content. • A positive correlation is found between the time spent looking at security indicators and the correct identification of phishing websites.
[21]	Two experiments, 60 and 45 participants	• Users who followed the authorization dialogue approach could identify permissions better than others. • The former group of users had a significantly higher average number of eye-gaze fixations on the permission text than other group participants.
[22]	Experiment, 50 participants	• The phishing susceptibility prediction model (DSM) had a higher correct prediction rate (92.34%) than that for individual feature prediction. • Combining static and dynamic features, DSM is an effective predictor of users' susceptibility to phishing.
[23]	Experiment, 4 participants	• All users focused on "Emergency Earthquake Warning." • Users with high literacy gazed at the domain name of the e-mail address.
[24]	Experiment, 23 participants	• When only checking content, phishing recognition performance returned an average error rate of 32.4% compared to 13.5% when security indicators are also checked. • The accuracy of user susceptibility to phishing based on eye movement is 79.3%.
[25]	Experiment, 107 participants	• In 26% of all cases, participants clicked on enclosed links or downloaded attachments in phishing e-mails. • In 38% of all cases, participants deleted or archived phishing e-mails in the spam folder, whereas they reported them in only 8% of all cases. • Experience and attention to security cues enable identifying and handling phishing e-mails. • Salient elements, such as logos, images or buttons divert attention from security cues more than plain text.
[26]	Experiment, 36 participants	• 90% of users depend on the domain name of a website as a legitimacy indicator. • Website design influences user decision on the legitimacy of a website.
[27]	Experiment, 40	• Knowledge and awareness about phishing were insufficient for

	participants	cyber protection, as even knowledgeable participants had insecure behaviours. • Attentiveness helps reduce susceptibility to phishing attacks. • Insecure behaviors continue to increase the likelihood of falling victim to phishing attacks.
[28]	Experiment, 22 participants	• Eye gaze fixation agreed with task performance. • Highlighted domains attracted visual attention, but did not effectively protect against phishing.
[29]	Experiment, 25 participants	• Users do not spend enough time analyzing key phishing indicators. • Longer fixations on login forms and logos may have regarded them as better than real legitimacy indicators. • Users who look longer at the login field are likely to have lower accuracy. • Personality traits (e.g, high attention control) improves phishing detection accuracy.
[30]	Usability study, 60 participants	• The domain name was used the most to determine legitimacy. • Less than 20% checked the SSL/TLS indicator. • Simple design is not necessarily better in mitigating phishing.
[31]	Simulated experiment, 41 participants	• Context-based micro-training increases user awareness. • Less than 10% of users could identify all phishing e-mails correctly. • Less than 50% of users evaluated all phishing identifiers.
[32]	Experiment, 132 participants	• Users unconsciously pay less attention to previously seen warnings. • Such habituation effect quickly sets in and progresses with successive warning exposures.

4. DISCUSSION

4.1. Technical Attributes

While scanning phishing materials, participants have shown to have some ability of recognizing some features associated to fraudulence. Yet, the general absence of a statistically significant correlation between detection accuracy and gaze fixation on the entire phishing material makes it unclear whether these materials, which exploit heuristics and invoke a cognitive miser style of processing, are successfully achieving their purpose. Nonetheless, participants rated e-mails containing misspelling rated as less trustworthy than others, as misspelling is a more categorical factor than urgency or threat indicators, which are open to personal interpretation.

For URLs, users can only expend a finite budget of resources to classify legitimacy. If the required resources exceed the budget, users will not expend them. Although threshold depends on factors other than the URL length, this notion is expected to apply generally.

Since fixation and dwell times are the highest for e-mail senders and website address bars, it is inferred that addresses are perceived as helpful phishing indicators. However, a single AOI does not necessarily translate into sound phishing determinations. This can be proven by the improved performance for users who studied multiple AOIs. Specific content, namely that asking for credit card information, was most likely to be identified as illegitimate. It can be assumed that most users have heard about phishing through the typical warning of messages asking for credit card information. In addition, contextual relevance and misplaced salience negatively impact security awareness. When

users face messages aligned with their work context, they pay less attention to security-related cues compared to when they are misaligned. As for salient design elements, they draw attention away from cues more than plain text.

4.2. Personal Attributes

Personal attributes, including experience, have shown to positively impact phishing detection. To demonstrate, users with past experience in web architecture, such as the ability to precisely interpret URLs, locks and page redirection, have demonstrated awareness by attending to a larger number of security-related information cues [35]. As such, it is inferred that experience allows the development of schemata [36] and identifying critical cues which enable pattern matching while forming awareness. On the other hand, other users seem to compensate for the lack of domain knowledge and experience by allowing more processing time to each e-mail or website. Comparing the average glance duration of both groups, it must be emphasized that the processing time for non-experts is a crucial factor.

An important factor which highly affects detection accuracy is attention control. Considered a personality trait [37], attention control has shown a high correlation with the ability to correctly detect phishing. It is characterized by pupil dilation [38], which provides an index of overall attentional effort, though it is time-locked to stimulus changes during attention.

Prior empirical studies have suggested mixed results on the association of personal attributes to phishing susceptibility. Yet, findings of this work agree with the majority of previous studies in terms of reporting insignificant correlations to agreeableness [39]. Therefore, a higher level of agreeableness does not translate to less attention to security-related cues.

Although domain knowledge, experience and attention control are key factors for mitigating phishing attacks [40], [41], it should be emphasized that they do not entirely guarantee user safety. Instead, risk awareness must be linked to a perceived vulnerability or a mitigation strategy, as perceived severity of consequences does not necessarily produce secure behavior.

4.3. Urgency and Threat Attributes

As for phishing indicators relating to urgency and threats, their immediate capture of human attention could be justified by survival information bias [42], in which humans prioritize processing information possibly related to their well-being. Security warnings have been shown to positively impact awareness by activating pattern matching mechanisms, which increase attentiveness to cues. Further, an indirect relationship between perceived threat and protection motivation suggests that perceiving phishing as threatening motivates users to take protective measures against phishing. Conversely, findings suggest that those who are aware of security-related cues are more confident in taking appropriate actions, thus exhibiting a higher perceived coping efficacy which positively influences protection motivation.

On the long term, the work in [32] presents a thorough study on habituation to visual stimuli, which demonstrates that habituation causes a steep decrease in attention after a few exposures. That is, it suggests that repeated exposure to security warnings may cause warnings to be physically seen, but not truly perceived by users. Specifically, gaze duration will decrease over successive viewings, and will decrease faster when viewing static warnings as compared to polymorphic warnings.

Taking these findings into account, it must be highlighted that computed detection accuracies in the studies are expected to be upper bounds on what users would achieve in practice, as additional safeguards in the artificial experimental setting would be removed.

5. IMPLICATIONS

Results demonstrate a number of fundamental points; they provide evidence that eye-tracking technology is useful in collecting gaze data on humans and phishing AOIs. Building upon this work provides more avenues for improving existing technology and increasing human awareness, all of which will be explored in this section.

5.1. User-Friendly URLs

From a technical standpoint, there exists no intrinsic security benefit to shortening a URL, beginning a domain name with www, or having a few special characters. In [20], although most users attempted at least occasionally to use the URL, they were not knowledgeable enough about URL structures to make informed decisions. For this reason, a more user-friendly URL bar should be developed. Specifically, domain names need to be more visually distinct to be effective security cues [43]. Alternatively, "breadcrumbs" could be used in browsers as in file managers to display the domain name more prominently, and users can view the whole URL by clicking on the URL bar. Since domain highlighting has not proven effectiveness, [16] recommends improving the design of indicators by changing the color, size or position to produce more salient cues for users.

5.2. Visual Aids for Browsing

The collected eye-tracking data can be useful in developing a gaze position indicator which informs the user when their gaze moves from one domain to another. That is, visualization could facilitate noticing changes even with less levels of attention. In [17], this data was used to develop a human-technical solution to guide user attention to the correct e-mail AOIs and therefore improve phishing detection accuracy. Nevertheless, the browser extension in [15] interacts with an eye-tracking device in order to develop satisfactory security behavior. By requiring users to look at the address bar before entering information, EyeBit checks whether users look at the address bar in browsers to improve security.

Real-time eye-gaze features can be used to automatically infer attentiveness states and assess the reliability of respective user response. Better yet, combining neural and ocular features will provide a robust detection system in which higher security measures will be achieved.

5.3. Embedded and Automated Authentication

Embedded authentication facilitates informing users about the legitimacy of a website. Yet, a potential implementation issue is the limited support for smartphones. Due to their constrained user interfaces by small screens, smartphone browsers often lack trustworthiness indicators. To solve this issue, it is recommended to implement a lightweight algorithm into smartphone browsers to deceptively detect phishing websites without user interaction. For instance, fake login credentials were used in [25] while simultaneously monitoring the destination server HTTP responses to authenticate a web page. Similarly, the UnPhishMe logic in [27] was implemented on a web browser to mitigate the exposure of login information to attackers, as well as eliminate zero-day and zero-hour phishing attacks in real-time.

5.4. Education and Training

The two most important factors in recognizing and, in return, mitigating phishing attacks are knowledge and attentiveness. At best, users should become experts to avoid falling victim for phishing. Through the assessment, comparison and improvement of training modules, training programs with heavy user involvement significantly impact user detection accuracy. For instance, educational materials and training strategies proposed by Merwe et al. in [42] compare phishing attacks with provided security service guidelines, and pinpoint weaknesses in the former if users adhere to the guidelines. Nonetheless, other training strategies were proposed in [44-46], and have proven effective in minimizing phishing susceptibility.

In [25], it was found that contextual relevance negatively impacts situational Information Security Awareness (ISA), which emphasizes the importance of tailoring phishing exercises to users and challenging employees with contextually relevant materials. On the other hand, training implementers must acknowledge the relevancy of each phishing material for trainees. To demonstrate, some employees may regularly interact with third-party groups, therefore increasing their exposure to phishing. In this case, they should acquire situational ISA by regularly matching AOI patterns with their mental library of what an AOI should look like to determine legitimacy.

To manage different abilities, difficulty levels of training sets should be personalized by varying the number of manipulated security cues to adhere to all trainees. One example is the implementation of EyeBit [15], which encouraged user attentiveness by tailoring training materials according to experience levels. For more experienced individuals, the variation of more difficult materials enhances their mental models and counters possible stereotypes which may have developed through repeated exposure. Conversely, less experienced individuals may benefit more from simpler materials of fewer manipulated cues to initially develop a mental library of prototypical phishing materials.

6. LIMITATIONS OF THE LITERATURE

As seen in Table 5, relatively small sample sizes were used in some studies. Compared to previous eye-tracking studies [47-49], this is not atypical. However, a small sample size is insufficient to investigate individual variability in how well eye-tracking estimates the ability to spot phishing attacks. Moreover, some eye-tracking data was excluded due to low validity scores arising from measurement devices and participant imprudence. For instance, sudden head, neck and/or face movements interfered with produced results. Consequently, reduced sample sizes may cause overfitting problems. Nevertheless, such problem can be suppressed by using head-mount eye-tracking devices or retaking measurements, if feasible. Also, some samples were of predominantly one gender. Although no evidence of gender differences in eye movements can be found [50], consistent research on the role of gender in phishing susceptibility remains a necessity [51]. Nonetheless, recruiting a more diverse sample and adopting the Bayesian optimization feedback loop [52], which adapts to unconsidered user groups, may clarify whether certain types of phishing are of more impact on different demographic groups.

The second limitation is the artificial setting of the experiments causing participation bias. Participants were explicitly informed that they are to spot phishing e-mails and/or websites. As demonstrated by [53], phishing detection accuracy may be higher when participants are aware in advance of the detection task. Therefore, results are expected to demonstrate an upper-bound on users' ability to correctly identify phishing, which is concerning given that detection accuracy was generally low. As such, a better experimental design would be to process phishing content as

a secondary task, where phishing content is randomly spread and participants are monitored to check if they would share sensitive data.

The third limitation is the utilization of online sources for the majority of phishing materials. Despite having an element of realism, some content was not ideal for experimental designs due to conflation of different phishing techniques, such as the combination of threat and urgency. Additionally, some content was presented to participants in the form of screenshots. While this method kept participants focused on the tasks, they were unable to interact with presented materials as they would in a real-life setting.

Another limitation is the classification of phishing content into AOIs, which may lead to correct detection but for completely wrong reasons. To modify this classification, AOIs could be formed only where phishing content can also be detected, and user perception as well as time taken to find these explicit recognition features can be studied. Moreover, given that some AOIs were relatively small, the error margin may have impacted the results, such as the recognition of the address bar in [15] by EyeBit. On the other hand, some used larger fonts and displayed URLs over multiple lines, which may have affected visual behavior and responses. A possible solution in this case is pattern matching in a digitized image, or estimating the position from the top-left browser corner. In both cases, it will be mandatory to adjust for each participant. Furthermore, it is to be noted that dwell time on an AOI does not necessarily reflect the level of understanding of a security cue. Conversely, a short glance duration does not necessarily indicate missing that element. In other words, it is possible to interpret the collected data differently.

Analyzing eye-tracking data is an objective step in the attempt to reflect and assess information security awareness. Although detection accuracy has generally improved compared to the past, it is unclear whether such improvement can be traced back to improved interfaces as opposed to increased user threat awareness. Considering these limitations, it is necessary to generate deeper and more valuable insights in order to form a comprehensive understanding and produce more results of high confidence.

7. CONCLUSIONS

In this work, 20 empirical studies have been pooled to examine phishing susceptibility through the lens of eye-tracking. Results provided empirical evidence that a user of a higher detection accuracy is characterized by higher attention control; spending more time looking at an AOI. Eye-tracking behavior has also shown that other attributes, namely technical background, domain knowledge, experience, training, and risk perception contribute to mitigating these attacks. In contrast, derived gaze position metrics, including TTFF, entry time and entry sequence data yielded inconclusive results regarding the impact on susceptibility to phishing attacks. It must be stressed that establishing user awareness has become of paramount importance, as one manipulated user could cause a catastrophic loss on both a personal and business infrastructural level. Thus, understanding how users determine the legitimacy of online content is a crucial step into developing usable security cues and training programs to mitigate phishing.

ACKNOWLEDGEMENTS

I would like to thank Dr. Shi Cao at the University of Waterloo for providing insight and knowledge into human factors testing, which steered me through this research. Dr. Shi Cao provided continuous encouragement, guidance and feedback and was always enthusiastic to assist in any way throughout the course.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.

[2] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

[3] Slandau, "Phishing Attack Statistics 2022," CyberTalk, 05-Apr-2022. [Online]. Available: https://www.cybertalk.org/2022/03/30/top-15-phishing-attack-statistics-and-they-might-scare-you/#:~:text=In%202021%2C%2083%25%20of%20organizations,s%20doubled%20since%20early%202020.

[4] M. Lukings and A. H. Lashkari, Understanding Cybersecurity Law and Digital Privacy: A Common Law Perspective. Cham: Springer International Publishing AG, 2022.

[5] R. Banerjee, Corporate Frauds Business Crimes Now Bigger, Broader, Bolder. SAGE Publications, 2022.

[6] "FBI: Internet Crime Report 2021," Internet Crime Complaint Center, pp. 22, 2022.

[7] 2021 Cyber Security Threat Trends: Phishing, Crypto Top the List. Cisco Umbrella. [Online]. Available: https://cloudmanaged.ca/wp-content/uploads/2021/09/2021-cyber-security-threat-trends-phishing-crypto-top-the-list.pdf.

[8] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond Eye Gaze: What Else Can Eyetracking Reveal about Cognition and Cognitive Development?," Developmental Cognitive Neuroscience, vol. 25, pp. 69–91, 2017.

[9] D. E. Hannula, "Worth a Glance: Using Eye Movements to Investigate the Cognitive Neuroscience of Memory," in Human Neuroscience, vol. 4, 2010.

[10] H. J. Green, P. Lemaire, and S. Dufau, "Eye Movement Correlates of Younger and Older Adults' Strategies for Complex Addition," Acta Psychologica, vol. 125, no. 3, pp. 257–278, 2007.

[11] K. Rayner, "Eye Movements in Reading and Information Processing: 20 Years of Research," Psychological Bulletin, vol. 124, no. 3, pp. 372–422, 1998.

[12] E. R. Grant and M. Spivey, "Eye Movements and Problem Solving: Guiding Attention Guides Thought," Psychological Science, vol. 14, pp. 462–466, Oct. 2003.

[13] B. Rehder and A. B. Hoffman, "Eyetracking and Selective Attention in Category Learning," Cognitive Psychology, vol. 51, no. 1, pp. 1–41, 2005.

[14] T. Whalen and K. Inkpen, "Gathering Evidence: Use of Visual Security Cues in Web Browsers," Proceedings of the Graphics Interface 2005 Conference, January 2005.

[15] J. McAlaney and P. J. Hills, "Understanding Phishing Email Processing and Perceived Trustworthiness Through Eye Tracking," Frontiers in Psychology, vol. 11, 2020.

[16] K. Pfeffel, P. Ulsamer, and N. H. Mu¨ller, "Where the User Does Look When Reading Phishing Mails– An Eye-Tracking Study," Learning and Collaboration Technologies. Designing Learning Experiences, pp. 277–287, 2019.

[17] D. Miyamoto, T. Iimura, G. Blanc, H. Tazaki, and Y. Kadobayashi, "EyeBit: Eye-Tracking Approach for Enforcing Phishing Prevention Habits," 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2014.

[18] L. Jaeger and A. Eckhardt, "Eyes Wide Open: The Role of Situational Information Security Awareness for Security-Related Behaviour," Information Systems Journal, vol. 31, no. 3, pp. 429–472, 2020.

[19] L. Huang, S. Jia, E. Balcetis, and Q. Zhu, "ADVERT: An Adaptive and Data-Driven Attention Enhancement Mechanism for Phishing Prevention," IEEE Transactions on Information Forensics and Security, pp. 1–1, 2022.

[20] N. Ramkumar, V. Kothari, C. Mills, R. Koppel, J. Blythe, S. Smith, and A. L. Kun, "Eyes on URLs: Relating Visual Behavior to Safety Decisions," ACM Symposium on Eye Tracking Research and Applications, 2020.

[21] "Eye-Tracking Phishing E-mails," Objective Experience SG Blog, 09-Nov-2017. [Online]. Available: https://eyetrackinginasia.wordpress.com/2017/11/09/eye-tracking-phishing-e-mails/.

[22] M. Alsharnouby, F. Alaca, and S. Chiasson, "Why Phishing Still Works: User Strategies for Combating Phishing Attacks," International Journal of Human-Computer Studies, vol. 82, pp. 69–82, 2015.

[23] Y. Javed and M. Shehab, "Look Before You Authorize: Using Eye-Tracking to Enforce User Attention towards Application Permissions," Proceedings on Privacy Enhancing Technologies, vol. 2017, no. 2, pp. 23–37, 2017.
[24] R. Yang, K. Zheng, B. Wu, C. Wu, and X. Wang, "Prediction of Phishing Susceptibility Based on a Combination of Static and Dynamic Features," Mathematical Problems in Engineering, vol. 2022, pp. 1–10, 2022.
[25] T. Matsuda, R. Ushigome, M. Sonoda, H. Satoh, T. Hanada, N. Kanahama, M. Eto, H. Ishikawa, K. Ikeda, and D. Katoh, "Investigation and User's Web Search Skill Evaluation for Eye and Mouse Movement in Phishing of Short Message," Advances in Intelligent Systems and Computing, pp. 131–136, 2019.
[26] D. Miyamoto, G. Blanc, and Y. Kadobayashi, "Eye Can Tell: On the Correlation Between Eye Movement and Phishing Identification," Neural Information Processing, pp. 223–232, 2015.
[27] L. Jäger and A. Eckhardt, "Phish Me If You Can: Insights from an Eye- Tracking Experiment," OPUS 4. [Online]. Available: http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/year/2021/docId/61421.
[28] A. Darwish and E. Bataineh, "Eye Tracking Analysis of Browser Security Indicators," 2012 International Conference on Computer Systems and Industrial Informatics, 2012.
[29] J. D. Ndibwile, E. T. Luhanga, D. Fall, D. Miyamoto, G. Blanc, and Y. Kadobayashi, "An Empirical Approach to Phishing Countermeasures Through Smart Glasses and Validation Agents," IEEE Access, vol. 7, pp. 130758–130771, 2019.
[30] A. Xiong, R. W. Proctor, W. Yang, and N. Li, "Is Domain Highlighting Actually Helpful in Identifying Phishing Web Pages?," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 59, no. 4, pp. 640–660, 2017.
[31] A. Neupane, M. L. Rahman, N. Saxena, and L. Hirshfield, "A Multi-Modal Neuro-Physiological Study of Phishing Detection and Malware Warnings," Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.
[32] A. Darwish and F. Aloul, "Impact of Page Design Factor on Cyber Security," 2014.
[33] J. Kävrestad, A. Hagberg, M. Nohlberg, J. Rambusch, R. Roos, and S. Furnell, "Evaluation of Contextual and Game-Based Training for Phishing Detection," Future Internet, vol. 14, no. 4, p. 104, 2022.
[34] B. B. Anderson, J. L. Jenkins, A. Vance, C. B. Kirwan, and D. Eargle, "Your Memory is Working Against You: How Eye Tracking and Memory Explain Habituation to Security Warnings," Decision Support Systems, vol. 92, pp. 3–13, 2016.
[35] "Phishing: Spot and Report Scam Emails, Texts, Websites and Calls," National Cyber Security Centre, 26-Nov-2021. [Online]. Available: https://www.ncsc.gov.uk/guidance/suspicious-email-actions.
[36] A. T. Duchowski, Eye Tracking Methodology: Theory and Practice. Cham: Springer, 2017.
[37] "The Role of a Schema in Psychology," The Role of a Schema in Psychology - Simply Psychology. [Online]. Available: https://www.simplypsychology.org/what-is-a-schema.html.
[38] J. S. Nairne, "Adaptive Memory: Evolutionary Constraints on Remembering," Psychology of Learning and Motivation, pp. 1–32, 2010.
[39] K. Kaspar and P. König, "Emotions and Personality Traits as High-Level Factors in Visual Attention: A Review," Frontiers in Human Neuroscience, vol. 6, 2012.
[40] T. Sommestad and H. Karlzén, "A Meta-Analysis of Field Experiments on Phishing Susceptibility," 2019 APWG Symposium on Electronic Crime Research (eCrime), 2019, pp. 1-14, doi: 10.1109/eCrime47957.2019.9037502.
[41] H. van Steenbergen, G. P. Band, and B. Hommel, "Threat but not Arousal Narrows Attention: Evidence from Pupil Dilation and Saccade Control," Frontiers in Psychology, vol. 2, 2011.
[42] R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006.
[43] D. Harley and A. Lee, "Phish Phodder: Is User Education Helping or Hindering?" in Proceedings of the Virus Bulletin Conference, 2007, pp. 1–7.
[44] A. Van der Merwe, M. Loock, and M. Dabrowski, "Characteristics and Responsibilities Involved in a Phishing Attack," in Proceedings of the 4th International Symposium on Information and Communication Technologies, Jan. 2005.
[45] M. Seckler, S. Heinz, S. Forde, A. N. Tuch, and K. Opwis, "Trust and Distrust on the Web: User Experiences and Website Characteristics," Computers in Human Behavior, vol. 45, pp. 39–50, 2015.

[46] O. A. Zielinska, R. Tembe, K. W. Hong, X. Ge, E. Murphy-Hill, and C. B. Mayhorn, "One Phish, Two Phish, How to Avoid the Internet Phish," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 58, no. 1, pp. 1466–1470, 2014.
[47] Z. A. Wen, Z. Lin, R. Chen, and E. Andersen, "What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game," Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.
[48] C. Nguyen, M. Jensen, and E. Day, "Learning Not to Take the Bait: A Longitudinal Examination of Digital Training Methods and Overlearning on Phishing Susceptibility," European Journal of Information Systems, pp. 1–25, 2021.
[49] J. J. Tecce, J. Gips, C. P. Olivieri, L. J. Pok, and M. R. Consiglio, "Eye Movement Control of Computer Functions," International Journal of Psychophysiology, vol. 29, no. 3, pp. 319–325, 1998.
[50] M. Libben and D. A. Titone, "Bilingual Lexical Access in Context: Evidence from Eye Movements During Reading," PsycEXTRA Dataset, 2007.
[51] W. Choi, M. W. Lowder, F. Ferreira, T. Y. Swaab, and J. M. Henderson, "Effects of Word Predictability and Preview Lexicality on Eye Movements During Reading: A Comparison Between Young and Older Adults," Psychology and Aging, vol. 32, no. 3, pp. 232–242, 2017.
[52] C. Klein, C. Klein, and U. Ettinger, Eye Movement Research: An Introduction to its Scientific Foundations and Applications. Cham, Switzerland: Springer, 2019.
[53] S. Kleitman, M. K. Law, and J. Kay, "It's the Deceiver and the Receiver: Individual Differences in Phishing Susceptibility and False Positives with Item Profiling," PLoS ONE, vol. 13, no. 10, 2018.
[54] F. Archetti, Bayesian Optimization and Data Science. Springer International Publishing, 2019.
[55] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "The Design of Phishing Studies: Challenges for Researchers," Computers Security, vol. 52, pp. 194–206, 2015.

AUTHORS

Noon Hussein was born in Edmonton, AB, Canada in 1999 to an immigrant Sudanese family. She received her BSc in Electrical Engineering from Qatar University in 2021, and is currently pursuing her MASc in Pattern Analysis and Machine Intelligence (PAMI) at the University of Waterloo, ON, Canada. Her research interests include IoT, systems design, cybersecurity and engineering education.

DEVELOPMENT OF A MONITORING SYSTEM FOR THE MANAGEMENT OF MEDICAL DEVICES

Kazuto Kakutani[1], Nobuhiro Ito[1], Kosuke Shima[1],
Shintaro Oyama[2] and Takanobu Otsuka[1]

[1]Nagoya Institute of Technology, Showa, Nagoya, Aichi, 466-8555, Japan
[2]Nagoya University, Chikusa, Nagoya, Aichi, 464-8601, Japan

ABSTRACT

In recent years, according to the sophistication of Medical Devices (MD), many portable MDs have been used and maintained with central management. However, the central management lends hospital staff the MDs only with managing by a ledger, therefore, missing or subletting may be caused. Furthermore, while the demand for the MDs is increasing due to the COVID-19, there is an issue that it is difficult to operate due to the shortage of clinical engineers against management duties of the MDs. In this study, we develop a power strip device which can measure electricity usage of plugged MD and its position and propose a visualization system for position and operation ratio of the MDs. We implemented 75 developed devices in three hospitals and confirmed that the system was effective to evaluate whether the number of the MDs owned by the hospital is appropriate.

KEYWORDS

Internet of Things (IoT), Wireless Network, Indoor Positioning, Medical Device, Management System

1. INTRODUCTION

In today's medical institutions, medical devices become smaller and more sophisticated. Many portable medical devices are used, and the number of devices requiring maintenance and management is increasing. Due to the 2007 revision of the Medical Care Law, the assignment of a medical device safety manager became mandatory, and the importance of medical device management increased in Japan [1-2]. In the case of a large hospital with about 1,000 beds, about 10,500 medical devices are managed, and approximately 90 devices are inspected per day. Because of this, the cost of managing the device is very high. Instead of managing different types of medical devices by department, a centralized management system centred on medical device managers such as medical engineers adopt in recent years [3-4].

The centralized management system manages the lending and returning of devices using a ledger. In this system, it is not possible to obtain the location and usage status of the rented device, but there are cases where the device is sub-leased or taken out temporarily without the lending process. As a result, they are left or lost in the course of medical work. The medical devices left for long periods of time cannot be regularly inspected. Searching for lost devices wastes medical device managers time and leads to lost opportunities.
In 2020, due to the prevalence of COVID-19, the demand for ventilators and extracorporeal membrane oxygenators used in intensive care units for critically ill patients surged. The Japanese government has requested an adequate supply of these medical devices. On the other hand,

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 87-102, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130407

securing enough medical devices leads to an increase in the number of managed and inspections device per day. It increases the burden on medical device managers. In order to reduce the burden of medical device management, it is necessary to obtain the location and operational status of medical devices. Localization and estimate Operation Status of them in hospitals has been widely studied.

One hospital introduced APM (Asset Performance Management) Service to manage device location and operation status. By attaching beacons and power monitors to medical devices and installing receiving stations in various places in the hospital, it is possible to obtain the operation status and location information of them. However, this service only informs that there is a medical device near the receiving station. Namely, it cannot specify the location. The operation status includes the state in which the device is on standby for prompt use to patients. Therefore, it is necessary to be able to measure the actual status of devices, including their maintenance cycle.
In our study, we focused on the fact that portable medical devices are usually transported on a medical wagon and used by connecting them to a power strip attached to the wagon. It supplies power to connected medical devices, measure the amount of current used, and transmit the obtained data to a server. These data are transmitted via LPWA (Low Power Wide Area) communication. LPWA communication is low power consumption and long-distance communication compared to Wi-Fi, which is already used in hospitals. Furthermore, security is safe because there is no Wi-Fi connection, and there is no influence of interference because the communication bandwidth is different. The system also transmits the RSSI (Received Signal Strength Indicator) of the radio waves transmitted by the Wi-Fi and BLE (Bluetooth Low Energy) beacons installed in the hospital. Using these data, we perform indoor position estimation in the hospital.

The estimated position is displayed on the hospital map for visualization. The operation status of medical device is estimated by collecting current consumption. In addition, we developed a system that visualizes the actual operation rate of each medical device category and evaluates the excess or deficiency of the number of medical devices owned.

2. RELATED WORKS

In the centralized management of medical devices, device information is digitized using barcodes to improve the efficiency of medical device lending and returning in general centralized management. This system enables the person in charge of device management, such as a medical engineer, to view logs of lending and returning operations. It contributes to cost reduction in medical device management by making it possible to view and search lending and returning logs. In addition, it improves the efficiency of lending and returning operations [4].

CMDMS (Computerized Medical Device Management System) has be proposed as an IT (Information Technology)-based medical device management system. It focuses on maintenance schedule management for the purpose of preventive and improved maintenance of medical devices. It demonstrated in an actual hospital [6]. This system manages device maintenance schedules based on the downtime and their frequency. These systems do not consider the real operation rate of the devices, such as the operation status or the measurement of the actual power consumption.

Indoor positioning has been studied very extensively, and positioning using image, infrared, ultrasound, Wi-Fi, RFID (Radio Frequency Identification), and even Bluetooth has been proposed [7]. In this paper, we focus on Wi-Fi, RFID, and Bluetooth, which do not have LoS (Line of Sight) constraints, assuming that the structure of a hospital is complex. Indoor

positioning system using medical devices equipped with wireless LAN (Local Area Network) has been constructed as a location estimation method [8]. This system performs positioning it by connecting them to Wi-Fi network in a hospital. The medical devices themselves must be capable of Wi-Fi communication, but modifications to the internal structure of existing medical devices, such as enabling Wi-Fi communication, are prohibited under the Pharmaceutical Affairs Law in Japan. Therefore, large-scale replacement is required to apply this system to all medical devices in a hospital. It requires a lot of labour and purchasing costs.

In indoor positioning, the positions of not only medical devices but also patients and medical staffs are estimated. A positioning system was constructed by attaching a bracelet to the patient's arm [9]. In this system, many BLE beacons are installed in the hospital. The bracelet receives the radio waves emitted from them, calculates the distance by measuring the strength of the radio waves, and performs positioning. On the other hand, the problem with this system is that it requires the installation of multiple beacons, gateways, and relay nodes in the hospital, making the system expensive to implement.

The above-mentioned indoor positioning using Wi-Fi and Bluetooth, it is important to conduct a radio wave strength survey at the location where the positioning is performed. Therefore, a system was proposed that tries Wi-Fi communication at each walking step to create fingerprints [10]. It is important guidelines for positioning by communicating Wi-Fi at comprehensive locations within a facility.

UHF (Ultra High Frequency) RFID is also used for indoor positioning because the tags attached to patients and medical devices do not require power [11]. RFID tag-based positioning has the advantage that the tags attached to the target are very inexpensive. However, unlike Wi-Fi-based positioning, which uses communication that has already been introduced, it requires the introduction of a new RFID reader, which increases the introduction cost [12].

On the other hand, indoor positioning using BLE beacons or RFID has the advantage. They can be reinforced after installation by adding their readers in areas where positioning is not possible due to lack of radio wave coverage.

For estimating operation status, a system and device was developed to monitor usage conditions using devices equipped with wireless LAN connectivity in recently years. Predicting future failures and performance of ventilators method has been proposed to optimize costs related to device inspections [13]. In this method, the data from maintenance and periodic inspections, including safety inspections of electrical characteristics and performance inspections, is used to determine whether it will pass inspections. The decision tree method has an accuracy of 98.5%. However, these systems require the replacement of devices to be managed, and the cost of implementation is high.

Recently, due to the prevalence of COVID-19, the development of ventilators as devices and the construction of systems that monitor them remotely [14]. A data model has been constructed for the purpose of improving the efficiency of ventilator management [15]. IT for medical devices has been studied. However, there is not much research on monitoring location information and operation information for medical devices in general.

In these methods, it becomes a new factor of a burden on medical staff such as battery replacement and inspection of IoT (Internet of Things) devices. In addition, the introduction and installation costs are high, and the burden at the time of introduction is heavy.

The current state of hospital management in Japan is that there is no budget to introduce these systems, and there are not enough human resources. Therefore, a system that can be introduced at a low cost and does not require much maintenance after the introduction is required.

3. PROPOSED SYSTEM ARCHITECTURE

Figure 1 shows the architecture of the system we propose in this paper. The IoT device we developed is connected to the power cable of a medical device. It collects current waveforms from medical devices. Wi-Fi access points and BLE beacons periodically transmit radio waves. Our device collects these radio waves. The collected current values and RSSI of these radio waves are transmitted by LoRa (Long Range) communication, which is a type of LPWA communication. The LPWA gateway receives data from IoT devices via LPWA communication. It transmits the received data to the DB (database) server. These collected data are accumulated in the DB server. The Web server acquires these data from the DB server and analyses it.

In the analysis, we estimate the indoor position in the hospital and the operation state of medical devices The analysis results are displayed on the website. Medical staffs view these data and use them for medical device management.

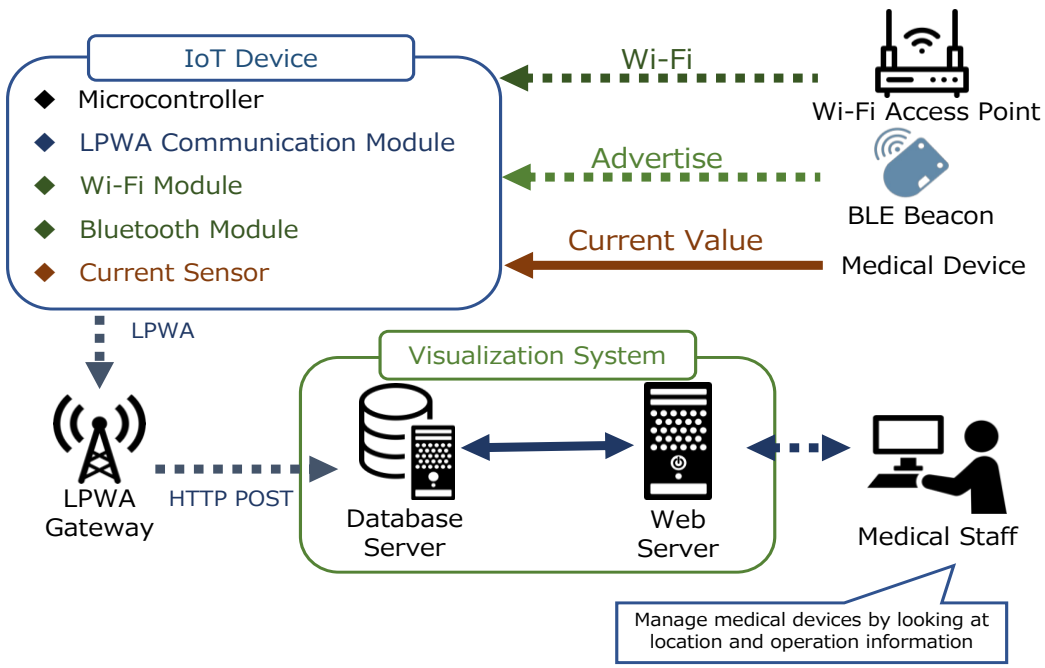


Figure 1. Proposed System Architecture

3.1 IoT Device for Medical Device

The configuration of the IoT device we designed and developed is shown in Figure 2. The device has the shape of a power strip with four outlets. It has an PCB (printed circuit board) inside. It has a MCU (Microcontroller Unit), LPWA Communication Module, Wi-Fi Module, Bluetooth Module, and four current sensors. These sensors measure the current value of each outlet.

The device is used as a power strip. The power cable of the medical device is connected to the outlet in our device. The power cable of our device is connected to the outlet of the hospital. Current is supplied to these medical devices through the current sensor. These current sensors

acquire the current and transmits it as analog data to MCU. The MCU converts the acquired value through an A/D (Analog to Digital) converter. The current value is calculated from these values. One current sensor is attached to each outlet. It is possible to acquire the current value of each medical device connected to the outlet.

To track the position of non-operating medical devices, our device is battery-powered and runs without a power supply from the power cable. When the power cable is connected, the AC(Alternate Current) -DC (Direct Current) converter of our device converts AC to DC and charges the litium ion battery.

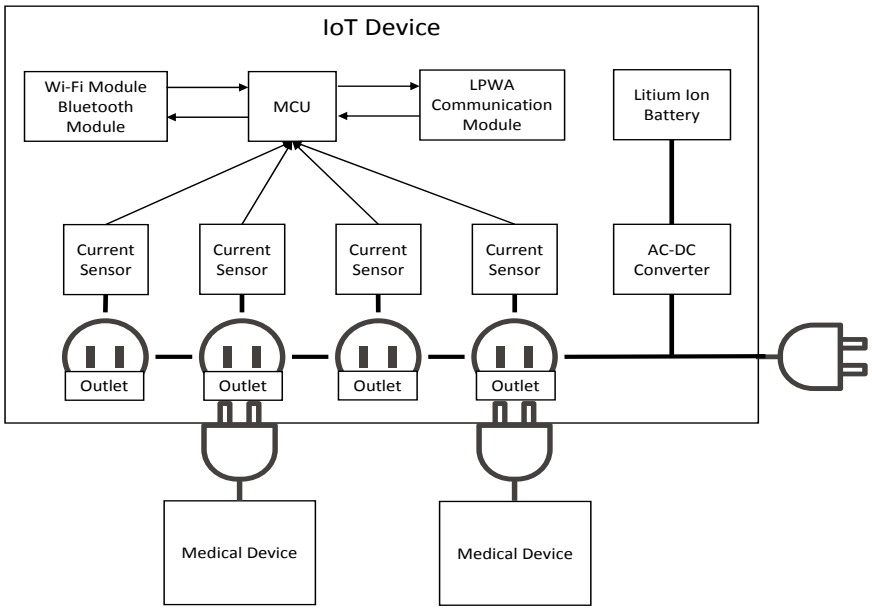


Figure 2. Configuration of IoT Device

3.2 Device Operation Flow

The flowchart of our device is shown in Figure 3. First, the current sensor acquired the AC current value. MCU calculates the RMS (Root Mean Square) of the current and the variance of the AC waveform from those values. When 10 minutes have passed since the operation and the current value and variance value are above the threshold, or when the timer has passed 12 hours, the current value and variance value are transmitted to the LPWA gateway via the LPWA communication module.

The Wi-Fi module and Bluetooth module acquire radio waves of Wi-Fi and BLE beacons. It transmits the acquired radio wave identifier and RSSI to the MCU. It transmits the collected radio wave data to the LPWA gateway via LPWA communication. Finally, reset the timer and repeat the process described above. By performing these processes, data for estimating the location information and operating status of the medical device used in the proposed system is collected.

3.3 Localization Method

We perform indoor location estimation using the RSSI of radio waves of Wi-Fi and BLE beacons (hereinafter referred to as nodes) collected by our device and their location coordinates. These coordinates (Geographic coordinate system) of the node are saved in advance in the DB of the web server. We focused on the fact that RSSI has the property that the farther the receiver (our

device) is from the transmitter (node), the weaker the signal strength. The position of the medical device is estimated by the triangulation method using the position of the node and the value of RSSI. This method has been widely studied and is commonly used [16-19].

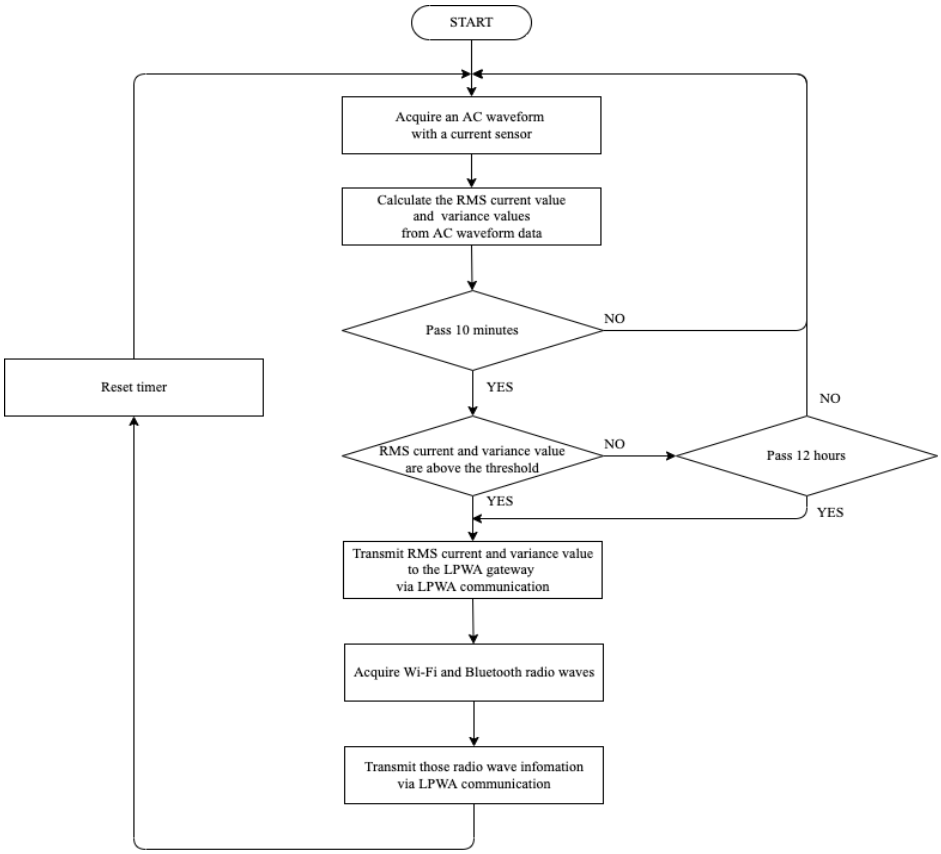


Figure 3. Flowchart of Device

However, since the radio wave characteristics are different between Wi-Fi and BLE beacons, there is a problem that the RSSI values differ greatly when measured from the same distance. Therefore, this method weights the RSSI of both Wi-Fi and BLE beacons. In this study, we conduct a demonstration experiment in hospitals targeting medical devices. In hospitals, Wi-Fi is installed in corridors, whereas medical device is rarely placed in that place. They are often used in a room, and it is important to estimate their location. In this system we propose, BLE beacons are installed in places where medical devices are used and stored, such as hospital rooms and storage areas without Wi-Fi access points. In addition, this system does not require high-precision positioning with an error of several centimetres as position estimation accuracy, and it is important to be able to estimate positions easily and inexpensively. As a result, increasing the RSSI weight for BLE beacons compared to Wi-Fi solves the above problem of having differences in RSSI between Wi-Fi and BLE Beacon.

The equation of the location estimation method considering the importance of Wi-Fi and BLE beacon information is shown below.

$$x = \frac{\sum_{k=1}^{n} w_i x_i}{\sum_{k=1}^{n} w_i} \#(1)$$

$$y = \frac{\sum_{k=1}^{n} w_i y_i}{\sum_{k=1}^{n} w_i} \#(2)$$

$$w_i = \begin{cases} W_{WiFi} \times 10^{\frac{RSSI_i}{10}} \text{ (Node is Wi} - \text{Fi)} \\ W_{BLE} \times 10^{\frac{RSSI_i}{10}} \text{ (Node is BLE Beacon)} \end{cases} \#(3)$$

$$-100 \leq RSSI_i \leq 0 \#(4)$$

Where x and y are the estimated location coordinates of the medical devices. x_i and y_i are the coordinates of node i in the geographic coordinate system. W_{WiFi} and W_{BLE} are the Wi-Fi and BLE beacon RSSI weights, and $RSSI_i$ is the RSSI of node i obtained by our device. (1) and (2) use the weights to calculate x (longitude) and y (latitude) of the device using the triangulation method. (3) converts the RSSI to power and multiplies the weights described above to calculate the weight of the acquired data of each node.

3.4 Device Operation Status Estimation Method

To estimate the operation status of medical devices, we use the RMS of current value and the variance of the AC waveform collected by our device. A medical device has an operation status during charging depending on whether it has an emergency battery, or it has multiple modes depending on the operation. The types and number of operation status to be estimated differ depending on the device. As an example, the operation status determination equation for a medical device that has three statuses, “Not Operating”, “Charging”, and “Operating”, is shown below.

$$\begin{aligned} 0 \leq I < \theta_{I1}, \quad & 0 \leq \sigma^2 < \theta_{\sigma^2 1} \Rightarrow Not\ Operating \\ \theta_{I1} \leq I < \theta_{I2}, \quad & \theta_{\sigma^2 1} \leq \sigma^2 < \theta_{\sigma^2 2} \Rightarrow Charging \\ \theta_{I2} \leq I \quad , \quad & \theta_{\sigma^2 1} \leq \sigma^2 \quad \Rightarrow Operating \end{aligned}$$

In this paper, "Not Operating" is the power cable of our device is not connected to the outlet in the hospital. "Charging" is the power cable of our device is connected to an outlet, the battery installed in the medical device is charging, and the medical device is not in operation. “ Operating” is the power cable of our device is connected to an outlet, and the medical device is in use.

I is the RMS of current value of the medical device calculated by the device. σ^2 is the variance value of the AC waveform. θ_n is the threshold of the RMS I and variance σ^2 value. The more operation status to be estimated for medical devices, the more thresholds are required for the determination.

The threshold is determined using the K-Means clustering method [20]. This method uses at least one week of data. The data are classified into 2 clusters by the K-Means clustering for each RMS of current value and variance value. A histogram of the collected data for one week for the medical device with the status of “Not Operating”, “Charging” and “Operating” is shown in Figure 4. Using K-Means clustering for the one-dimensional data of the current value and the variance value, we performed two-state classification. As a result, by setting 400 as the threshold for the current value and 5000 as the threshold for the variance value, it was shown that it is possible to distinguish between the two statues.

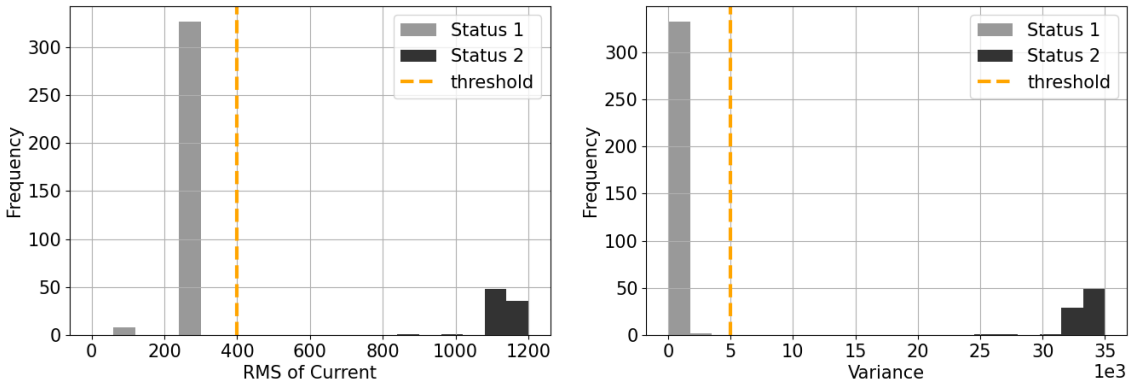


Figure 4. Histogram of current and variance values. It shows all the data. Thresholds are indicated by dashed lines. The RMS of current threshold is 400. The Variance threshold is 5000.

Next, K-Means clustering is performed on these data by dividing the intervals into two groups: those with a variance greater than or equal to 5000 and those with a variance less than or equal to 5000.

The histograms with variance values below 5000 are shown in Figure 5. We perform two-state clustering on these data. As a result, by setting 150 as the threshold for the current value and 500 as the threshold for the variance value, it is possible to distinguish between the two statuses.

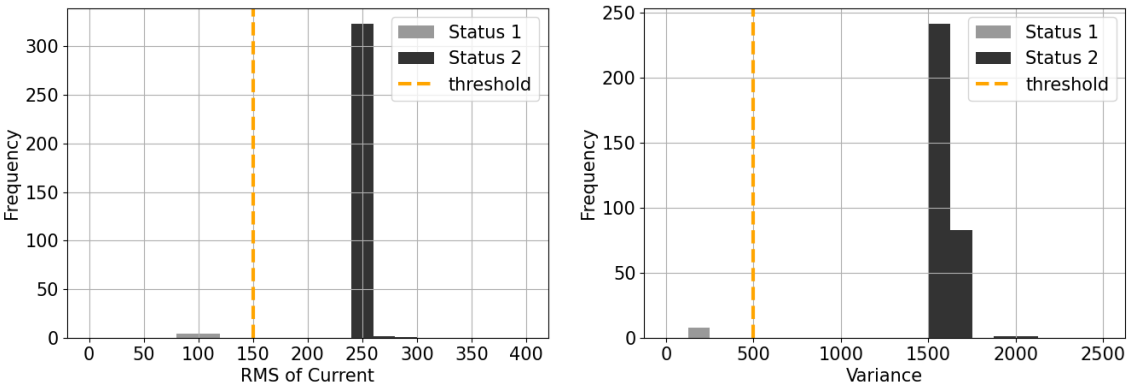


Figure 5. Histogram of current and variance values. It shows data with variance value less than 5000. Thresholds are indicated by dashed lines. The RMS of current threshold is 150. The Variance threshold is 500.

The histograms with variance values above 5000 are shown in Figure 6. We perform two-state clustering on these data. As a result, by setting 1050 as the threshold for the current value and 30000 as the threshold for the variance value, it was shown that it is possible to distinguish between the two statuses. In this identification, there is only one data set with a current value below 1050 and a variance value below 30000, which is classified as one state. We regard this state as noise from the current sensor and call it the undefined status.

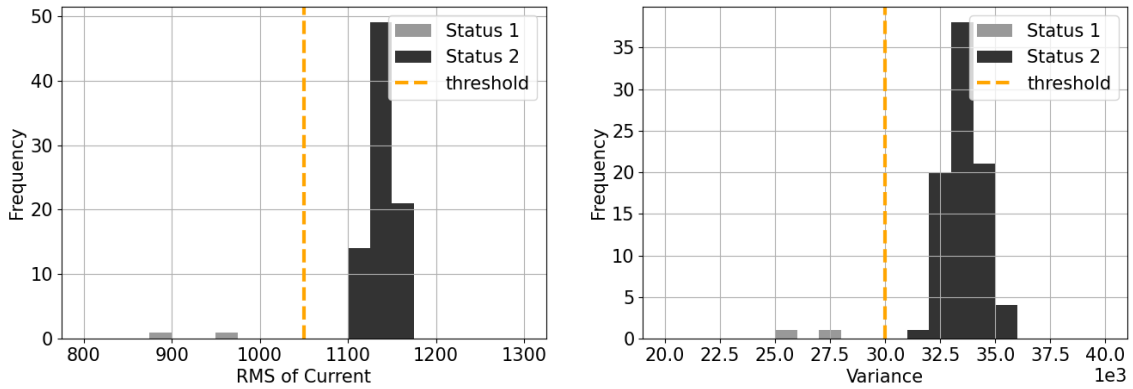


Figure 6. Histogram of current and variance values. It shows data with variance value greater than or equal to 5000. Thresholds are indicated by dashed lines. The RMS of current threshold is 400. The Variance threshold is 5000.

Finally, the following equation is defined.

$$\begin{aligned} 0 \le I < 150, \quad & 0 \le \sigma^2 < 500 && \Rightarrow Not\ Operating \\ 150 \le I < 300, \quad & 500 \le \sigma^2 < 5000 && \Rightarrow Charging \\ 300 \le I < 1050, \quad & 5000 \le \sigma^2 < 30000 && \Rightarrow Undefined \\ 1050 \le I \quad , \quad & 30000 \le \sigma^2 && \Rightarrow Operating \end{aligned}$$

I is the RMS of current value of the medical device calculated by the device. σ^2 is the variance value of the AC waveform. θ_n is the threshold of the RMS I and variance σ^2 value.
The more operation status to be estimated for medical devices, the more thresholds are required for the determination. Undefined status means that the current and variance values are calculated unexpectedly, and if these values appear frequently, we define a new status. These processes are performed periodically and evaluated.

3.5 Visualization Web Site

The visualization page viewed by medical staffs displays the details, location, operation information, and operation rate of medical devices. Visualization page is shown in Figure 7. On the map screen in the centre, the hospital map of the floor specified by the tab at the top is displayed.

Medical staffs can check the location and operation status of medical devices by the icon shown on the map. In the medical device list at the bottom of the map screen, the information and operation rate of the devices on the map are displayed. Detailed information about medical devices can be displayed by selecting the device from the list. The detailed information includes location information, operation information, medical device names, and management numbers within the hospital.

Medical staffs can use this visualization page to check the location and the current operation information and operation rate of the medical devices. Therefore, this system can be used to search for lost medical devices, confirm device that should be used with priority. Furthermore, by using the collected data, we will use it to optimize the number of owned medical devices.

Figure 7. Visualization Web Page. Circle icons in the hospital map indicate the location of medical devices. The color of the icon indicates the status of the device.

4. DEMONSTRATION EXPERIMENT

Experiments were conducted at the following three hospitals in Aichi and Hyogo prefectures, Japan.

- University Hospital A with about 1000 beds (in Aichi Prefecture)
- Civilian Hospital B with about 220 beds (in Aichi Prefecture)
- University Hospital C with about 1000 beds (in Hyogo Prefecture)

Figure 8 shows how the developed device is attached to the medical device. The number of medical devices with the development device installed is shown in Table 1. The most ventilator has a heating humidifier for humidifying the air, which is connected to the outlet of the developed device, but some ventilator connects a heated humidifier to the ventilator itself. There are 6 types of ventilators, 10 types of ultrasound diagnostic equipment, and 2 types of hemodynamic monitoring equipment. Although the ventilator and hemodynamic monitoring devices are equipped with a battery, there are some devices that have the battery removed.

Table 1. List of medical devices that have introduced our system

Hospital	Medical Device	Auxiliary Equipment	Number
Hospital A	Ventilator	Heated Humidifier	36
Hospital A	Hemodynamic Monitoring Device	None	14
Hospital B	Ventilator	Heated Humidifier	2
Hospital B	Ultrasonic Diagnostic Equipment	None	12
Hospital B	Electrocardiograph	None	1
Hospital C	Ventilator	None	10

Figure 8. Our device attached to a medical device

5. RESULTS AND ANALYSIS

Regarding the location information of medical devices, we introduced our system at a hospital where we are conducting a demonstration experiment and evaluated it. The experiment has been conducted since November 14, 2021.

5.1 Evaluation of Localization

At the target hospital, a beacon is installed in each room, and position estimation is performed using radio waves from Wi-Fi and BLE beacons. It was well-received that the accuracy was sufficient for ordinary tasks such as searching for devices.

At Hospital A, a medical engineer provided the following events that demonstrate the usefulness of this system. "With the existing medical device management system, it was not possible to identify the current location of the medical device. I was able to locate it easily by using the system."

As a result, we obtained the opinion that our system is useful. In the future, we will evaluate the usefulness of this system by comparing it with other highly accurate position estimation systems.

5.2 Evaluation of Operation Status Estimation

In an interview with a medical engineer, we received the opinion that hospital A may have an excessive number of hemodynamic monitoring devices and ventilators before the introduction of this system. Therefore, we focused on the number of hemodynamic monitoring devices in operation each day and evaluated the appropriate number of devices. The data acquisition period is from January 18, 2022 to March 24, 2022. The daily number graphs of two types of hemodynamic monitoring devices are shown in Figure 11 and Figure 12. In the figures, the solid line indicates the total number of units in operation for the day and the dashed line indicates the average number of units in operation during the acquisition period.

Hospital A owns 5 hemodynamic monitoring devices A shown in Figure 9. The number of devices operating simultaneously on February 15 was 5 units. It is the maximum number of devices being used. The average number of devices in operation was 1.8 units. Thus, we thought that the number of 5 units of hemodynamic monitor A was appropriate. The medical engineer commented that it would be possible to acquire data over a longer period and determine whether to add several devices.

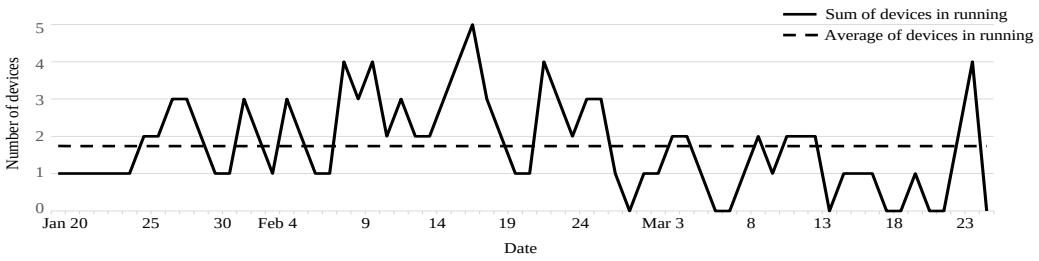


Figure 9. Number of Operating Hemodynamic Monitoring Device A

Hospital A owns the number of 9 units of hemodynamic monitoring devices B shown in Figure 10. During the data acquisition period, the maximum number of devices in operation was 2 units, and the average number of devices in operation was 0.9 units. It is considered that the number of 9 units of hemodynamic monitoring device B owned during this period is excessive for actual operation. Therefore, Hospital A was able to decide to operate fewer devices B.

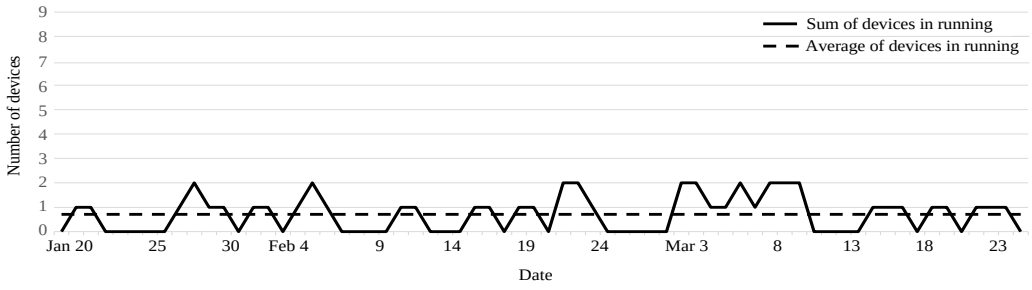


Figure 10. Number of Operating Hemodynamic Monitoring Device B

The system we propose can show real-time operation status and operating rate data of medical devices as a basis for overs and shorts of medical devices in hospitals. It also can propose the optimization of the number of owned units.

6. CONCLUSIONS

In this paper, we developed a medical device management system that supports the management of various medical devices owned by medical institutions, and conducted a demonstration experiment. The IoT device we developed estimates the location information and operation status of medical devices. Our device functions as a power strip, and it can be introduced to various medical devices to collect data without violating medical device approval in Japan.

When the system is introduced, location estimation is performed only with existing Wi-Fi. BLE beacons are installed as necessary in places where detailed location estimation is required. The system we proposed can be operated with only one LPWA gateway installed per target area. For these reasons, it can be introduced inexpensively and easily.

In the demonstration experiment, with the cooperation of three hospitals, the device we developed was introduced for 75 medical devices and evaluated in an actual environment. We interviewed medical engineers and evaluated the system. As a result, we showed that the accuracy of location estimation is sufficient for practical use. We have also shown that the judgment of operation status is useful as data for effectively the management of medical devices, such as optimizing the number of devices.

In the future, we will miniaturize the device to reduce costs and introduce it into small portable medical equipment. We compare our system with a high-precision positioning system and evaluate the estimation accuracy quantitatively. We improve the location estimation accuracy by considering the time series of data and optimizing the BLE beacon placement. In the operating state estimation, a machine learning method is used, and the device infers the operating state on the edge side, so that more states will be automatically judged. Furthermore, it detects anomalies such as equipment failure. In cooperation with the medical device management system, the operating status is evaluated.

ACKNOWLEDGEMENTS

This work was partially funded by the Strategic Information and Communications R&D Promotion Programme (SCOPE) grant number 215006007.

REFERENCES

[1] Ministry of Health, Labour and Welfare, "Law enforcement for amendmend On Partial Enactment of a Law for Partial Amendment of the Medical Act to Establish a System for Providing Quality of Health Care (full text in Japanese)," Japanese Society for Quality and Safety in Healthcare, Vol. 2, No. 2, pp. 190-196, DOI: 10.11397/jsqsh.2.190 (2007)

[2] H. Atarashi, M. Hirose, H. Ide, and S. Koike, "The current status and issues of medical equipment safety managers 10 years after deployment and the role of clinical engineers," The Japanese journal of medical instrumentation, Vol. 90, No. 3, pp. 245‐255, DOI: 10.4286/jjmi.90.245 (2020).

[3] M. Hirose, T. Kano, H. Atarashi, F. Aoki, T. Takakura, et al, "Current Status and Future Issues in Safety Management of Medical Equipment," Vol. 16, No. 1, pp. 43 ‐ 50, DOI: 10.11397/jsqsh.16.43(2021).

[4] Y. Tamaki, Y. Kitahara, S. Mitsuda, R. Kishino, and M. Tsunekawa, et al, "Effect of central management system of medical equipment by introducing ME equipment lending system," The Japanese journal of medical instrumentation, Vol. 90, No. 4, pp. 363‐369, DOI: 10.4286/jjmi.90.363 (2020).

[5] R. Hosaka, "Wireless Communication Technology in the present Medical Scene," Shonan Institute of Technology Journal, Vol. 55, No. 1, 2021.

[6] A. Nirapai, A. Wongkamhang, M. Sangworasil, R. Saosuwan, P. Yotthuan, et al, "Computerized Medical Device Management System," 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-ICON), pp. 1–4 (online), DOI: 10.1109/TIMES- iCON.2018.8621812 (2018).

[7] L. Mainetti, L. Patrono, and I. Sergi, "A survey on in-door positioning systems," 2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 111–120 (online), DOI: 10.1109/SOFTCOM.2014.7039067 (2014).

[8] N. Kurashima, A. Okubo, H. Seshima, A Kozakai, Y. Sato, et al, "Comparison of location tracking accuracy of Ⅳ infusion & Syringe pumps with wireless LAN and active tag(Part 1)," The Japanese journal of medical instrumentation, Vol. 88, No. 1, pp.2-8, DOI: 10.4286/jjmi.88.2 (2018)

[9] G. Shipkovenski, T. Kalushkov, E. Petkov, and V. Angelov. "A Beacon-Based Indoor Positioning System for Location Tracking of Patients in a Hospital," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6 (online), DOI: 10.1109/HORA49412.2020.9152857 (2020).

[10] T. -Y. Chang and Y. -R. Chien, "Indoor Positioning Method for Smart Mobile Device Based on Fuzzy Wi-Fi Fingerprint," 2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), 2020, pp. 1-2, doi: 10.1109/ICCE-Taiwan49838.2020.9258035.
[11] T. Akahane, R. Hosaka, and T. Murohashi, "A Study of Management System for Medical Equipment using UHF Passive RFID Tags," The Japanese journal of medical instrumentation, Vol. 33, No. 1, pp. 15-25, DOI: 10.14948/jami.33.15 (2013).
[12] H. Atarashi, K. Tanaka, and H. Tamai, "Development and evaluation of a WiFi-based location detection system for medical equipment management," The Japanese journal of medical instrumentation, Vol. 79, No. 6, pp. 373-381, DOI: 10.4286/jjmi.79.373 (2009).
[13] Ž. Kovačević, L. Gurbeta Pokvić, L. Spahić, and A. Badnjević, "Prediction of medical device performance using machine learning techniques: infant incubator case study," Health and Technology, Vol. 10, pp. 151-155, DOI: 10.1007/s12553-019-00386-5 (2020).
[14] M. N. Mohammed, H. Syamsudin, M. A. H. Abdelgnei, D. Subramaniam, M. A. A. M. Taib, et al, "Toward a Novel Design for Mechanical Ventilator System to Support Novel Coronavirus (Covid-19) Infected Patients Using IoT Based Technology," 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS), 2021, pp. 294-298, doi: 10.1109/I2CACIS52118.2021.9495900.
[15] N. Mori, K. Sekine, and T. Takakura, "Development of a data model for ventilator operation and maintenance," Vol. 91, No. 4, pp. 342-349, DOI: 10.4286/jjmi.91.342 (2021).
[16] D. Qian and W. Dargie, "Evaluation of the reliability of RSSI for indoor localization," 2012 International Conference on Wireless Communications in Underground and Confined Areas, 2012, pp. 1-6, doi: 10.1109/ICWCUCA.2012.6402492.
[17] S. L. Shue and J. M. Conrad, "Development of a portable X Bee C library and RSSI triangulation localization framework," 2014 11th Annual High Capacity Optical Networks and Emerging/Enabling Technologies (Photonics for Energy), 2014, pp. 125-128, doi: 10.1109/HONET.2014.7029375.
[18] M. Sakai and H. Morita, "Indoor Location Estimation using Bluetooth Devices," The Japan Society for Management Information (JASMIN), National research presentation, 2016, pp. 53-56.
[19] T. -H. Lee, S. Weng and J. Sanford, "Indoor radio triangulation using only RSSI data," 2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting, 2020, pp. 1097-1098, doi: 10.1109/IEEECONF35879.2020.9329964.
[20] J. A. Hartigan, and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108, 1979.

AUTHORS

Kazuto Kakutani received the B.S. degree in the Department of Computer Science of Nagoya Institute of Technology in March 2021. Since April 2021, he has been with the Computer Science Program, Graduate School of Engineering, Nagoya Institute of Technology, Japan

Nobuhiro Ito has been with the Department of Computer Science, Nagoya Institute of Technology, Japan

Kosuke Shima is Assistant Professor of Nagoya Institute of Technology. He received the M.E. and Doctor of Engineering degrees from the Nagoya Institute of Technology in 2016 and 2021.His main research interests include pattern recognition, physical motion analysis, non-supervised learning algorithms, and anomaly detection.

Shintaro Oyama is an Associate Professor at the Innovative Research Center for Preventive Medical Engineering (PME), Nagoya University, Tokai National Higher Education and Research System. Since 2020, he has been involved in the development of medical AI and medical xR technologies at the Medical xR Center and the Hub for Medical Health Data Integration Research and Education.

Takanobu Otsuka is Associate Professor of Nagoya Institute of Technology. He received the M.E and Doctor of Engineering degrees from the Nagoya Institute of Technology in 2011 and 2016. From 2012 to 2015, he was an Assistant Professor of the Nagoya Institute of Technology. From 2015 and 2016, he was a visiting researcher at UCI (University of California Irvine). His main research interests include IoT, Multi-agent systems, intelligent agents, distributed system, and software engineering for off shoring.

SALES FORECASTING OF PERISHABLE PRODUCTS: A CASE STUDY OF A PERISHABLE ORANGE DRINK

T. Musora, Z. Chazuka, A. Jaison, J. Mapurisa, and J. Kamusha

School of Natural Sciences and Mathematics,Department of Mathematics,
Chinhoyi University of Technology ,Pivate Bag-7724, Chinhoyi, Zimbabwe

ABSTRACT

The primary goal of any organization involved in trading business is to maximize profits while keeping costs to a bare minimum. Sales forecasting is an inexpensive way to achieve the aforementioned goal. Sales forecasting frequently leads to improved customer service, lower product returns, lower deadstock, and efficient production planning. Because of short shelf life of food products and importance of product quality, which is of concern to human health, successful sales forecasting systems are critical for the food industry. The ARIMA model is used to forecast sales of a perishable orange drink in this paper. The methodology is applied successfully. ARIMA $(0,1,1)(0,1,1)_{12}$ was concluded as the appropriate model. Model diagnostics were done; results showed that no model assumption was violated. Fitted values were regressed against observed values. A very strong linear relationship was evident with an R^2 value of over 90% which is very plausible.

KEYWORDS

Sales forecasting, Orange Drink, ARIMA, Model Diagnostics, R^2- value.

1. INTRODUCTION

Forecasting plays a key role in decision-making and business planning. Probably, the most important function of business is forecasting. Demand forecasting in brief is an estimation of a supply chain constituent's expected sales in a specified future period [1] A forecast is a starting point in planning. The objective of forecasting is to minimize the risk in decision making. To a large extent, success depends on getting those forecasts rightly, [2] gives some important forecasting applications for the strategic areas in business. Also, [2] explains the types of forecasts and managerial planning. From this explanation, it can be concluded that a distribution company aims to determine the optimal supply of orange drinks that minimizes costs or maximizes profits in the face of uncertain demand. In the case of more shipping than demand the company has undue costs caused by stocking, high returns, transportation, and other operational issues, or in the case of less shipping than demand the company has sales lost. [3] used machine learning to forecast horticultural sales and concluded that machine learning outperforms classical forecasting on horticultural sales. Classical forecasting methods for example Autoregressive Integrated Moving Average and Exponential Smoothing are nevertheless widely used in research and industry. Regardless of their rather simple concept, they often show a competitive performance. ([4],[5] , [6]).

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT 2023
pp. 103-116, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130408

[7] states that reliable forecasts are essential for a company to survive and grow. In a manufacturing environment, management must forecast the future demands for its products and on this basis provide for the materials, labor, and capacity to fulfill these needs. These resources are planned and scheduled well before the demands for the products are placed on the firm. Forecasting is the heart and blood of any inventory control system. A firm with hundreds or thousands of items must anticipate in advance demands that will occur against these items. This is needed to have the proper inventory available to fill customers' demands as they come in. Management must plan several months for this inventory since procurement lead times from suppliers generally runs from one to six months. With each time, forecasts are needed for the months in the planning horizon. The forecasts are used to determine whether or not an order to the supplier is needed now and if so how large the order should be [7] explains that forecasting techniques can be categorized into three groups. The first is called qualitative, where all information and judgment relating to an item are used to forecast the item's demands. This technique is often used when little or no demand history is available. The forecasts may be based on marketing research studies, the Delphi method, or similar methods. The second group is called causal, where a cause-and-effect type of relation is sought. Here, the forecaster seeks a relation between an item's demands and other factors, such as business industrial, and national indices. The relationship is used to forecast the future demands of the item. The third group is called time series analysis, where a statistical analysis of past demands is used to generate the forecasts. A basic assumption is that the underlying trends of the past will continue into the future. This paper is primarily concerned with forecasting as it relates to time series analysis. In this context, the time series represents the demands recorded over past time intervals. The forecasts are estimates of the demands over future time intervals and are generated using the flow of demands from the past. This paper proceeds as follows. Section 2 gives the literature review, some theoretical structures for exponential smoothing models, and autoregressive integrated moving average (ARIMA) models. Section 3 includes comprehensive empirical results and analysis of orange drink circulation and results. Section 4 is the discussion and conclusion

2. TIME SERIES ANALYSIS AND MODELLING STRATEGY

The importance of predicting future values of a time series cuts across a range of disciplines. Economic and business time series are typically characterized by trend, cycle, seasonal, and random components. Powerful methods have been developed to capture these components by specifying and estimating statistical models. These methods comprise; log transformation, square root transformation exponential smoothing, and ARIMA, which are described by [9] and [10]. They reveal that ARIMA gives more accurate out-of-sample forecasts on average compared to other smoothing methods, although ARIMA requires much more effort. [11] states that exponential smoothing originated in Robert G. Brown's work as an OR analyst for the US Navy during World War II. [12] identify that the more sophisticated exponential smoothing methods seek to isolate trends or seasonality from irregular variation. Where such patterns are found, the more advanced methods identify and model these patterns. The models can then incorporate those patterns into the forecast. Exponential smoothing uses weighted averages of past observations for forecasting. The effect of past observations is expected to decline exponentially over time. [13] states that the exponential smoothing methods are relatively simple but robust approaches to forecasting. They are widely used in business for forecasting demand for inventories. Three basic variations of exponential smoothing are given simple exponential smoothing, trend-corrected exponential smoothing, and the Holt-Winters method. [14] states that the ARIMA method developed by [15] is one of the most noted models for time series data prediction and is often used in econometric research. The ARIMA method has been originated from the autoregressive (AR) model, the moving average (MA) model, and the combination of the AR and MA, the ARMA model. Compared with the early AR, MA, and ARMA models, the ARIMA model is

more flexible in application and more accurate in the quality of the simulative or predictive results. [15] highlight that in the ARIMA analysis, an identified underlying process is generated based on observations to a time series for generating a good model which shows the process-generating mechanism precisely.

[17] and [18] have considered that the only problem with ARIMA appears that the modeling is mathematically sophisticated in theory and requires a deep knowledge of the method. Therefore, building an ARIMA model is often a difficult task for the user, requiring training in statistical analysis, a good knowledge of the field of application, and the availability of an easy-to-use but a versatile specialized computer program. The BoxJenkins approach to modeling and forecasting time series data is but one of a large family of quantitative forecasting methods which have been developed in the fields 12 of operations research, statistics, and management science. Box-Jenkins models are also known as ”ARIMA” models, the acronym standing for Autoregressive Integrated Moving Average. This terminology is made clear in the following sections. Exponential smoothing, linear regression, Bayesian forecasting, and generalized adaptive filtering are some of the other techniques which are termed ”extrapolative” forecasting [6]. Many of these methods have a common element; they utilize only the previous values of a series of numbers to forecast the future values of interest. Hence, they are referred to as univariate models, since the values from a single variable are used to predict the future values of the same variable. This is in contrast to multivariate models, where the variable of interest is also considered to depend on other variables

2.1. ARIMA Model

The ARIMA model is an extension of the ARMA modelling the sense that by including auto-regression and moving average it has an extra function for differencing the time series. If a dataset exhibits long-term variations such as trends, seasonality and cyclic components, differencing a dataset in ARIMA allows the model to deal with them. Two common processes of ARIMA for identifying patterns in time-series data and forecasting are auto-regression and moving average.

2.2. Autoregressive Process

Most time series consist of elements that are serially dependent in the sense that one can estimate a coefficient or a set of coefficients that describe consecutive elements of the series from specific, time-lagged (previous) elements. Each observation of the time series is made up of random error components (random shock; a_t) and a linear combination of prior observations.

2.3. Moving Average Process

Independent from the autoregressive process, each element in the series can also be affected by the past errors (or random shock) that cannot be accounted for by the autoregressive component. Each observation of the time series is made up of a random error component (random shock, ϵ) and a linear combination of prior random shocks.

2.4. Autoregressive Integrated Moving Average Process, ARIMA (p,d,q)

A series $X_t$is called an autoregressive integrated moving average process of orders p,d,q, ARIMA(p,d,q), if $W_t= \nabla^d X_t$, where $W_t$is the differenced time series.

We may define the difference operator ∇ as $\nabla X_t = X_t - X_{t-1}$. Differencing a time series $\{X_t\}$ of length n produces a new time series $\{W_t\} = \{\nabla^d X_t\}$ of length $n-d$. If $\{Z_t\}$ is a purely random process with mean zero and variance σ_z^2, the general autoregressive integrated moving average process is of the form

$$Wt = \phi 1 Wt-1 + \phi 2 Wt-2 + ... + \phi p Wt-p + Zt + \theta 1 Zt-1 + ... + \theta q Zt-q$$

In terms of the backward shift operator, the ARIMA(p,d,q) process is

$$\Phi_p(B) W_t = \Theta_q(B) Z_t$$

Remark: The autoregressive integrated moving average process is specifically for non-stationary time series. The differencing transformation is useful in reducing a nonstationary time series to a stationary one.

2.5. Seasonal Auto-regressive Integrated Moving Average Process

Let s, be the number of observations per season. Then the time series, X_t, is called a seasonal autoregressive integrated moving average process of orders p,d,q, seasonal orders P,D,Q and seasonal period s, if it satisfies;

$$\phi_p(B)\Phi_P(B^s)\nabla^d \nabla_s^D X_t = \theta_q(B)\Theta_Q(B^s) Z_t$$

Where $\nabla_s^D X_t = \sum_{j=0}^{D} \binom{D}{j} X_{t-js}$, and $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of order p and$q$, that is ;

$$\phi_q(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right)$$
$$\theta_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

We identified the stationary component of a data set by performing the Ljung and Box test. We tested this hypothesis by choosing a level of significance for the model adequacy and compared the computed Chi-square (χ^2) values with the (χ^2) values obtained from the table. If the calculated value is less than the actual(χ^2) value, then the model is adequate, otherwise not. The $Q(r)$ statistic is calculated by thefollowing formula:

$$Q_(r)) = n(n+2)\sum \frac{r^2(j)}{n-j}$$

where$n$ is the number of observations in the series and $r(j)$ is the estimated correlation at lag j. Furthermore, we tested the data to specify the order of the regular and seasonal autoregressive and moving average polynomials necessary to adequately represent the time series model. For this purpose, model parameters were estimated using a maximum likelihood algorithm that minimized the sums of squared residuals and maximized the likelihood (probability) of the observed series. The maximum likelihood estimation is generally the preferred least square technique. The major tools used in the identification phase are plots of the series, correlograms (plots of autocorrelation and partial autocorrelation verses lag) of the autocorrelation function (ACF) and the partial autocorrelation function (PACF).The ACF and the PACF are the most important elements of time series analysis and forecasting. The ACF measures the amount of linear dependence between observations in a time series that are separated by a lag k. The PACF plot helps to determine how many autoregressive terms are necessary to reveal one or more of the

following characteristics:time lags where high correlations appear, seasonality of the series, and trend either in the mean level or in the variance of the series. In diagnostic checking, the residuals from the fitted model were examined against their adequacy. This is usually done by correlation analysis through the residual ACF plots and by goodness-of-fit test using means of Chi-square statistics. At the forecasting stage, the estimated parameters were used to calculate new values of the time series with their confidence intervals for the predicted values.

2.6. Performance Valuation

To choose the best model among the class of plausible model, the estimated parameters were tested for their validity using, ACF , PACF, Probability Plot and Histogram of residuals, a time series plot of observed and fitted values and other error statistics such as coefficient of determination(R^2) were analysed.

2.7. Data Source

The data used in this research is historical data of monthly sales of cases of the perishable drink from a small drink manufacturing company in Harare, Zimbabwe which among other products manufactures the perishable orange drink. Each case contains 24 bottles of the drink. The company intends to minimise losses due to returns of the drinks as result of reduced shelf life.

3. RESULTS AND ANALYSIS

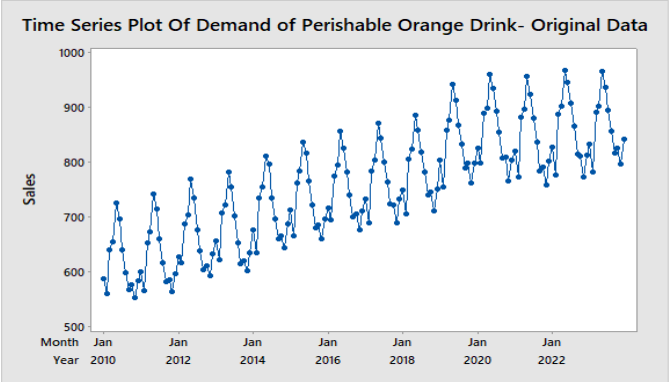


Fig.1: Time Series Plot Of Demand of Perishable Orange Drink- Original Data

Visual inspection of the plot shows that the series is dynamic. So need is there to transform the data so as to make it stationary.

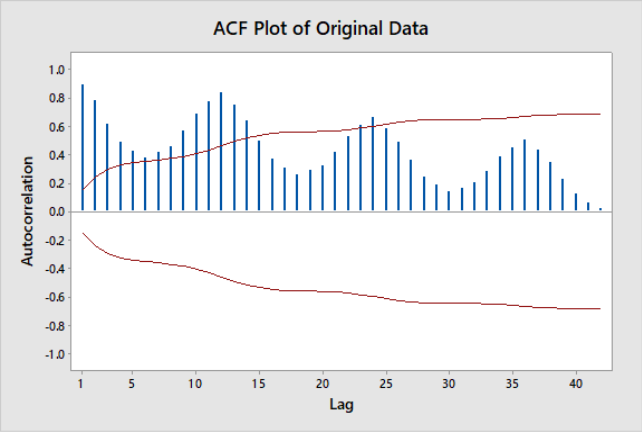


Fig.2: ACF Plot of Original Data

ACF of most lags are very high, there is evidence of positive and negative autocorrelation. This is a typical ACF plot of a non stationery time series. Thus a model cannot be fitted at this stage. This further affirms need to transform the data.

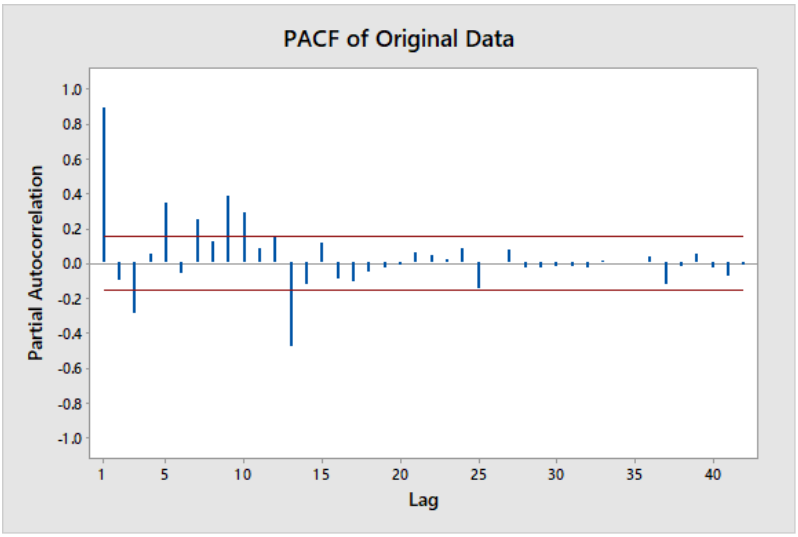


Fig.3: PACF plot of Original Data

The PACF plot shows a number of significant spikes, which is typical of a non stationary series. Thus we have to transform the data to make it stationary.

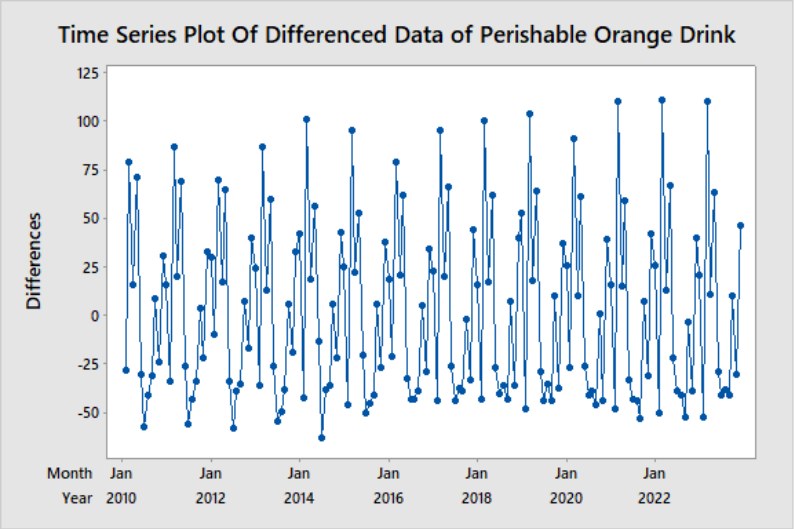


Fig.4: Time series plot of Differenced Data of Perishable Orange Drink

Visual inspection of the plot reveals that the differenced series fluctuates around zero, thus the data is now stationary

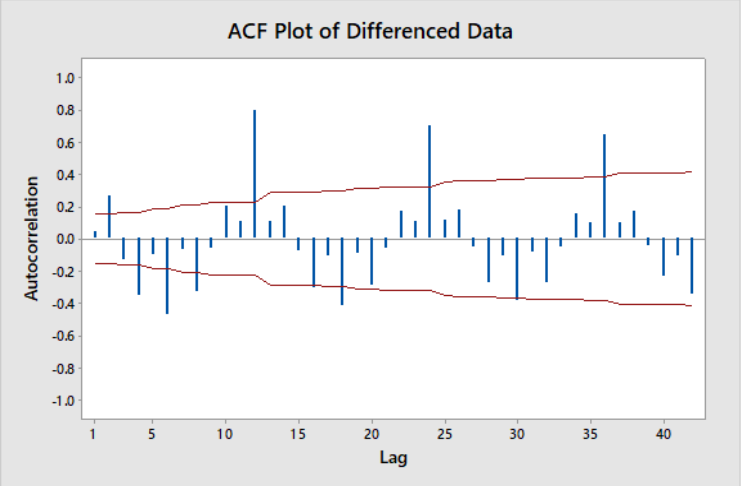


Fig.5: ACF plot of Differenced Data

The ACF shows a significant spike at lag 2 and there is evidence of negative dumped oscillations with the rest of the ACF's essentially zero, hence a seasonal ARIMA model is suggested

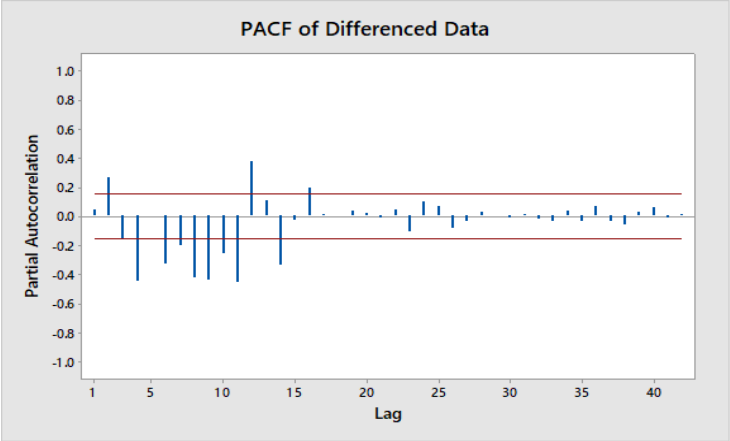


Fig.6: PACF plot of Differenced Data

PACF plot shows a significant spike at lag 2 which is seasonal and there is evidence of negative dumped oscillations with the rest of the PACFs essentially zero, hence a seasonal ARIMA model is also suggested.

3.1. Parameter Estimation

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
MA 1	0.9707	0.0321	30.21	0.000
SMA 12	0.6533	0.0660	9.90	0.000
Constant	-0.00181	0.01501	-0.12	0.904

Differencing: 1 regular, 1 seasonal of order 12 Number of observations: Original series 167, after differencing 154 Residuals: *SS* = 8667.31 (back forecasts excluded) *MS* = 57.40
DF = 151 Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	99.2	232.2	326.0	405.3
DF	9	21	33	45
P-Value	0.000	0.000	0.000	0.000

Thus the fitted model is SARIMA$(0,1,1)(0,1,1)_{12}$

3.2. Model Diagnostics

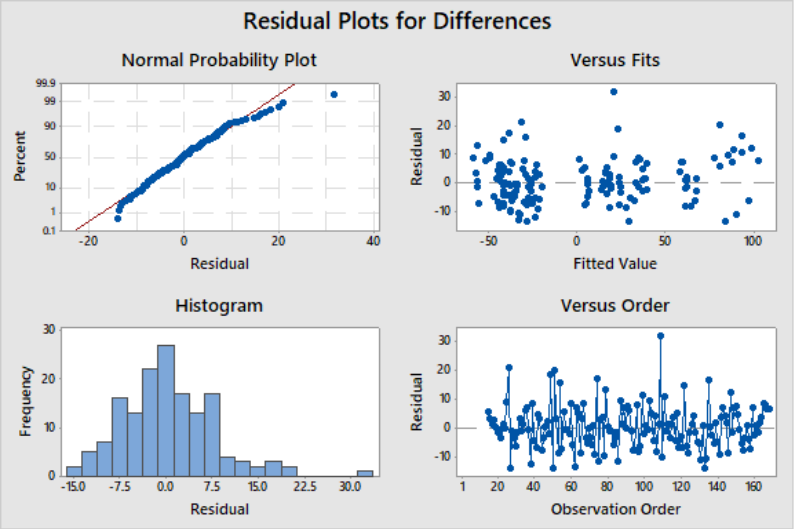


Fig.7: Residual Plot for Differences

The normal probability plot is almost a straight line, an indication that the normality assumption has not been violated. A plot of residuals against fitted values shows no pattern and the histogram of residuals also indicates that the normality assumption has not been violated. Hence the fitted model is good and thus can be used for forecasting.

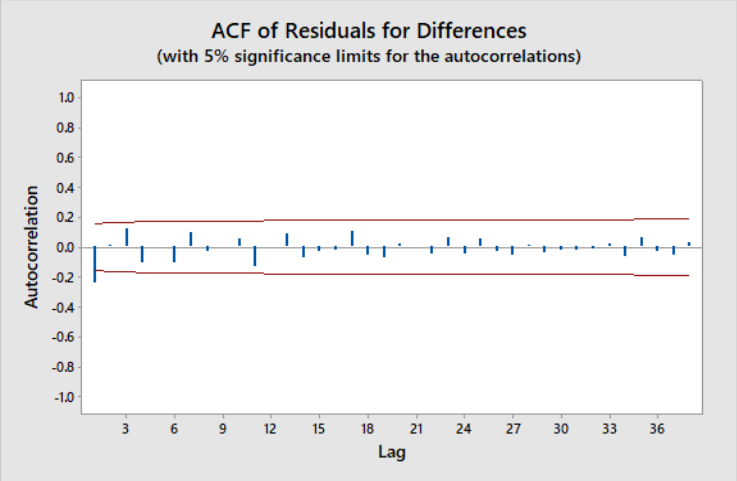


Fig.8: ACF of residuals for Differences

Figure 8 ACF plot has no significant spikes suggesting that there might be no possible additional parameters which may have been omitted in this model.

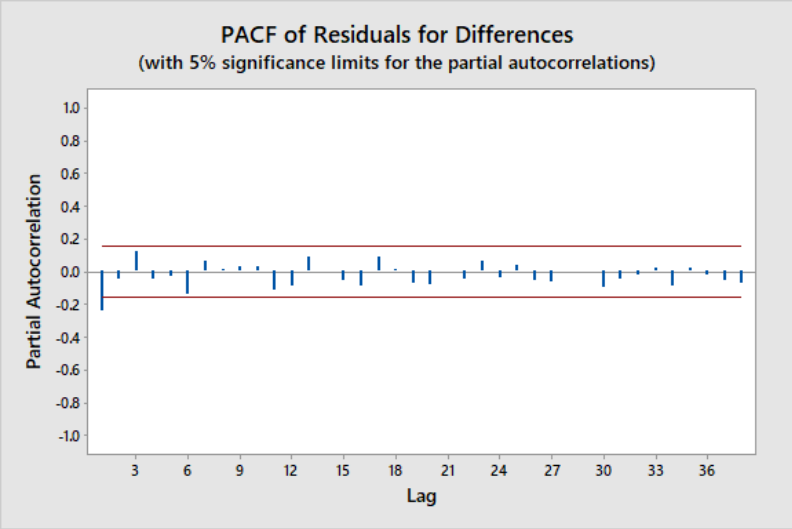


Fig.9: PACF of residuals for Differences

Figure 9, The PACF plot of residuals refuses any significant spikes suggesting that there might be no possible additional parameters that may have been omitted in this model. Since the fitted model appears good enough, it can be used for forecasting future demand of the perishable orange drink.

3.3. Inference Based on the Model

3.3.1. Forecasts from period 159

Period	Forecast	Lower	Upper	Actual
159	847.49	783.94	911.04	892.00
160	836.74	752.93	920.55	903.00
161	896.10	795.37	996.84	966.00
162	897.13	782.01	1012.25	937.00
163	881.36	753.45	1009.26	896.00
164	869.28	729.77	1008.80	858.00
165	831.16	683.39	983.87	817.00
166	808.89	670.91	991.40	827.00
167	826.39	639.32	978.55	797.00
168	828.86	639.23	1013.47	
169	830.30	609.70	1024.07	
170	831.37	591.24	1050.89	
171	833.43	573.95	1072.51	
172	834.99	557.99	1092.21	
173	836.55	543.07	1111.99	
174	838.11	529.03	1130.03	
175	839.67	515.73	1147.19	
176	841.23	503.09	1163.60	
177	842.79	491.03	1179.36	
178	842.79	479.47	1194.55	
179	844.35	487.93	1209.23	
180	845.91	468.36	1223.45	
181	847.47	457.67	1236.26	

The fitted values compares well with the observed values, thus the fitted model is reliable.

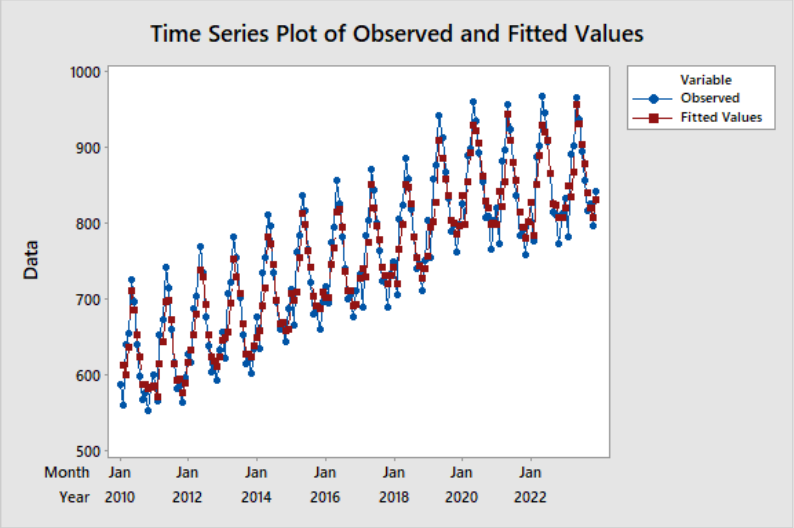


Fig.10: Time Series plot of Observed and Fitted Values

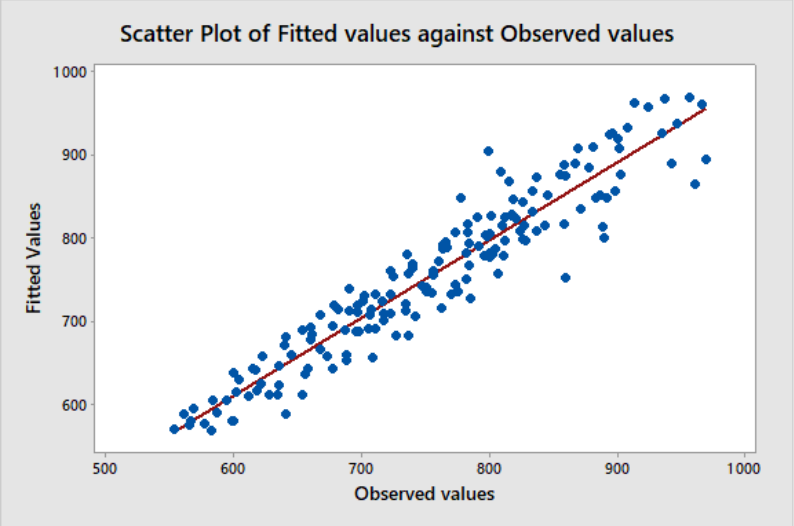


Fig.11: Scatter plot of Fitted values against Observed Values

3.4. Regression Analysis: Fitted Values versus Sales

The scatter plot of fitted values against observed values suggests a positive linear relationship.

Method

Rows unused 1

Analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	150339	153391	1500.63	0.000
Error	165	165303	1002		
Total	166	1668694			

Model Summary

R- sq	R-sq(adj)	R-sq(pred)
90.09%	90.03%	89.84%

Coefficients

Term	Sales	Se Coef	T-Value	
Constant	48.4	18.4	2.62	0.009
Sales	0.9360	0.0242	38.74	0.000

Regression Equation

$$\text{Fitted Values} = 48.4 + 0.9360 \times \text{Sales}$$

The coefficient of determination value is 90.09% indicates that the fitted model accounts for about 91% of the variation in the fitted values. Thus the fitted seasonal ARIMA model which generated the fitted values must me appropriate and hence can be used to forecast sales values.

4. DISCUSSIONS AND CONCLUSIONS

This study demonstrates how ARIMA time series and Regression models are useful to study and forecast sales for a particular company. This paper demonstrates also how the Time Series Forecasting System can be used to construct a model of forecasting. The ARIMA$(0,1,1)(0,1,1)_{12}$ predicted the data considerably well and gave reliable forecasts. According to the data presented, this model was best in forecasting the sales, but could not tell why the sales will contain outliers. The Time Series forecasting system helped construct a model, the ARIMA time series and the Regression, which is effective for forecasting and can be applied to other businesses in order to plan their sales. However, it would be interesting to do further research on the factors that influence the sales, such as the growth of the population of consumers, the industrial growth in the region, the immigration, and so on; this would consolidate better this company's planning.

REFERENCES

[1] Islek, I.; Oguducu, S, 2017. A Decision Support System for Demand Forecasting based on ClassifierEnsemble. In: Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, 13: 2017, p. 35–41.

[2] Shim, J.K. (2009). Strategic Business Forecasting: Including Business Forecasting Tools and Applications, Global Professional Publishing.

[3] Haselbeck, F. Killinger, J, Menard, K, Hannus T and Grim , G.D.(2020) Machine Learning OutperformsClassical Forecasting on Horticultural Sales Predictions

[4] Kolassa, S. (2021). Commentary on the M5 forecasting competition. International Journal of Forecasting, http://dx.doi.org/10.1016/j.ijforecast.2021.08.006, Advance online publication

[5] Makridakis, S., Spiliotis, E., &Assimakopoulos, V. (2020). The M4 competition: 100, 000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74. http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.
[6] Makridakis, S., Spiliotis, E., &Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. International Journal of Forecasting,
[7] http://dx.doi.org/10.1016/j.ijforecast.2021.07.007, Advance online publication
[8] Thomopoulos, N.T. (1980). Applied Forecasting Methods, Prentice Hall.
[9] Wei, W.W.S. (2006). Time Series Analysis: Univariate and Multivariate Methods, Second edition, Addison Wesley.
[10] Granger, C.W.J. and Newbold, P. (1974). Spurious regressions in econometrics, Journal of Econometrics,2, pp. 111–120.
[11] Reid, D.J. (1975). A review of short-term projection techniques, Practical Aspects of Forecasting, H.Gordon, ed., London: Operational Research Society.
[12] Gass, S.I. and Harris, C.M. (2000). Encyclopedia of operations research and management science,Centennial edition, Dordrecht, The Netherlands: Kluwer.
[13] Yafee, R. and McGee, M. (2000). Introduction to Time series Analysis and Forecasting with Applicationof SAS and SPSS, Academic Press.
[14] Gardner, E.S. (1985).Exponential smoothing: The state of the art, Journal of Forecasting, 4, pp. 1–28.
[15] Ediger, V.S., Akar, S. and Ugurlu, B. (2006). Forecasting production of fossil fuel sources in Turkeyusing a comparative regression and ARIMA model, Energy Policy, 34, pp. 3836–3846.
[16] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, Revised ed.Holden-Day, San Francisco, USA.
[17] Gardner, E.S. (2006). Exponential smoothing: The state of the art – Part II, International Journal ofForecasting, 22, pp. 637–666.
[18] Ho, S.L. and Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis, Computers Industrial Engineering, 35, pp. 213–216.
[19] Melard, G. and Pasteels, J.M. (2000). Automatic ARIMA modeling including interventions, using timeseries expert software, International Journal of Forecasting, 16, pp. 497–508.
[20] Billah, B., King, M.L., Snyder, R.D. and Koehler, A.D. (2006). Exponential smoothing model selectionfor forecasting, International Journal of Forecasting, 22, pp. 239–247.
[21] Cho, S.H. and Song, I. (1996). A Formula for Computing the Autocorrelations of the AR Process, TheJournal of the Acoustical Society of Korea, 15, 4–7.
[22] Diebold, F.X. (2001). Elements of Forecasting, Second edition, Thomson Learning.
[23] Enders, W. (1995). Applied Econometric Time Series, John Wiley and Sons, New York.
[24] Eshel, G. (2003). The Yule Walker Equations for the AR Coefficients, Citeulikearticle- id: 763363.
[25] Fuller, W.A. (1996). Introduction to Statistical Time Series, Second edition, John Wiley and Sons,New York.

AUTHORS

T Musora Is a PhD student at Chinhoyi University of Technology, He is currently being supervised by Dr. F Matarise from the university of Zimbabwe.He holds a MSc in Operations Research from NUST, Zimbabwe, B.Sc., (Special Hons)in Operation Research & Stats from NUST, Zimbabwe, BSc., Maths & Stats with Zimbabwe Open University, Zimbabwe and has a Dip. Ed.from Gweru Teachers College, Zimbabwe. Mr Musora is also a lecturer in the Department of Mathematics at Chinhoyi University of Technology.

Dr. Z Chazuka Holds a PhD in Mathematical Biology from University of South Africa. She also holds an Msc In operations Research from the National University of Science of Technology Zimbabwe, and a B.Sc Applied Mathematics from the National University of Science of Technology . Currently, she is lecturer at Chinhoyi University of Technology. Her research interests is in Mathematical Biology.

J Mapurisa received a MSc In Mathematics from University of Zimbabwe. He holds a B.Sc in Mathematics from the University of Zimbabwe . Currently, he is Lecturer at Chinhoyi University of Technology. His research interests is in Fluid Dynamics (flow in channels) .

A Jaison is a Lecturer in the Department of Mathematics at Chinhoyi University of Technology. He holds a MSc in Operations Research from NUST, Zimbabwe, B.Sc. in Operation Research from NUST, Zimbabwe. His research interest is in Multivariate Analysis, Financial and Statistical Modeling

J Kamusha received aMsc In Mathematical Sciences from Stellenbosch University. He holds a Bachelor's degree in Mathematics majoring in Actuarial Science . Currently, he is Lecturer at Chinhoyi University of Technology. His research interests is in Graph Theory and Convolutional Neural Networks .

ANALYZING AND PERSONALIZING THE LEARNING PERFORMANCE FOR SPECIAL NEEDS STUDENTS USINGMACHINE LEARNING AND DATA ANALYTICS

Eric Xiong[1], Yu Sun[2]

[1]Crean Lutheran High school, 12500 Sand Canyon Ave, Irvine, CA 92618
[2]California State Polytechnic University, Pomona,
CA, 91768, Irvine, CA 92620

ABSTRACT

We design a server-client system that collects students' engagement information and reports it to a centralized server to help teachers assist neurodivergent students in order to provide a visual representation of students' engagement status aiming to promote an equal learning opportunity for neurodivergent students [6].

In recent years, everyone throughout the globe are all seeking higher education, either for themselves, or for their children. Students are learning an increasing amount in classes and have needed to spend a lot more effort and attention to succeed. In this race for higher education, a specific group of underrepresented minorities has been left behind. This group being the neurodivergent population, specifically high-functioning people with ASD(Autism Spectrum Disorder) [7]. These students often require more attention due to hypersensitivity, and a shorter attention span than the neurotypical populace. These students have all that's necessary to learn and understand the material, although teachers are often stuck to a faster pace curriculum that does not easily allot so much attention to a singular student. Due to this problem many teachers believe that a efficient way to passively gauge these students attentiveness would significantly benefit their education. This paper develops a server-client system that collects students' engagement information and reports it to a centralized server to help teachers assist neurodivergent students in order to provide a visual representation of students' engagement status aiming to promote an equal learning opportunity for neurodivergent students. We applied our application to [Class] and conducted an Evaluation of the approach based on the qualitative data collected from the students.

KEYWORDS

Facial features, information collection, Education

1. INTRODUCTION

1.1. Engagement Detection

What did this paper do, what does it contribute, and why did we not choose this one (what does it do, their conclusions,why we didn't use it.)

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 117-125, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130409

1.1.1. Paper 1

This paper attempts to use a students facial expression, head position, and eye gaze to calculate a students engagement level using computer vision and machine learning. The results from their algorithm showed to be 10% from their baseline. We did not select this model to continue because we believe we could find something with higher accuracy, or can be directed towards a more general case, because everyone disengages in different ways, especially for members of the neurodivergent community.

1.1.2. Paper 2

The paper uses the features of the subjects face such as eye gaze, and head pose using OpenFace to gauge the engagement levels of the subject [8]. The results from their program claims to have a 90% accuracy. We decided we did not select this model because the algorithm is too complex, since many students, especially neurodivergent ones, have different ways of disengagement leading to a more simplistic model being more general and effective.

1.1.3. Paper 3

This paper reviewed many methods of determining student engagement levels during the course of an educational environment. The paper concludes that although promensing the computer vision process of determining engagement is still bound by many limitations. We did not select this model because it does not actually present a model of its own instead reviewing previous and established methods presenting their benefits and flaws.

1.2. Neurodivergent Students Education

The neurodivergent, specifically people with autism, typically show certain characteristics that can be considered as a hindrance in mainstream education [9]. Some of these traits include difficulty focusing, hyperactivity and unpredictable mood changes. Many establishments have already attempted to accommodate such needs, but there is still a widespread demand for methods and aids to fit aforementioned accommodations.

1.2.1. Project 1

Special attention is typically needed for teachers to assist neurodivergent students, although many times, especially in standard classroom settings design for neurotypical students, the teachers are not equipped with the necessary methods or tools to assist the students [1][10]. This has led to an increasing demand for teachers to be better trained about students with autism both in education and behavior management.

1.2.2. Paper 2

PRT as a a method, and how it is helpful. Pivotal Response training(PRT) is claimed as a behavioral treatment of Autism based on the principles of Applied Behavioral Analysis(ABA) [2]. This process is initiated by the child and typically involves the use of games to aid in the process. This process attempts to develop communication and language skills, increase positive social behaviors and provide relief from disruptive self-stimulatory behaviors.

Our tool can helps PRT in terms of social intervention through the process of real-time providing student’s engagement info

There have been intensive studies on detecting human emotion and engagement status through facial expressions cite(1,2), where the paper uses … … … to detect. It provides a general overall interface but yet it lacks usability due to a missing interface and updating system. A secondary problem is that such algorithms only detect a single user at a time.

Server-client based engagement detection dashboard integrated with xxx's paper on engagement detection [3].

1. real-time update on dashboard
2. class-student hierarchies
3. adjustable algorithm and devices
4. alert when disengaged

In this paper, we will server-client based engagement detection dashboard integrated with xxx's paper on engagement detection. Our goal is to keep the students in a more attentive state for a longer period of time. This will aid students in their education while also helping teachers make their classes more fruitful and entertaining by understanding when the students are disengaged. Some Useful features of our tool are that it updates on a dashboard in real-time, class-student hierarchies, adjustable algorithms and devices, and alert when disengaged. Therefore, we believe that the tool would allows teachers to better manage their classes of neurodivergent students.

In two application scenarios, we demonstrate how the above combination of techniques increases student engagement and teacher efficiency. First, we show the usefulness of our approach by a comprehensive case study on the evolution of student engagement. This will be accomplished by using the algorithm on Multiple classes of different students and then asking for their engagement directly, with a reward to encourage honesty. Once these values are obtained we will calculate the percent error of the detected result from the alleged result to measure the accuracy of the algorithm. Second, we analyze the teacher feedback using the System usability scale. This will be accomplished by surveying the teachers after use of the system. This will measure the helpfulness, effectiveness and quality of the system.

1. student's engagement accuracy validation

use algorithm on students for a day, measure their average engagement, and at end of class, ask is the average correctly reflected their engagement status

2. teacher engagement assisting

how much time does it save you in terms of monitoring engagement with/without this tool Introduction of the background, open problem, solution and special contribution, and paper structure The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Understanding Neurodivergent Student's Attention Span

Our first challenge was understanding Neurodivergent student's attention span is inherently difficult because of its diversity and differences with the general neurotypical population. This is because every case of Neurodivergence is different both in severity and in symptoms. Often Students may have other issues that also hinder attentiveness such as ADHD(Attention deficit/hyperactivity disorder) [11]. This also makes attentiveness incredibly difficult to track because of its individual nature. Most neurodivergent students disengage in different waysleading to the tracking of disengagement also difficult. Most solutions are also not universal and only work on the people it was specifically designed for. Many people have overcome this issue with their own methods to become famous.

2.2. Designing A Online Learning System that Reflects Students' Engagement Status in Real- Time

Our second challenge was designing an online learning system that reflects students' engagement status in real-time [12]. Real time was significant for our program because it is needed for this system to achieve its fundamental purpose. If the system did not update the engagement of the students in real time, it would not be possible for the program to inform the teacher that some of the students are disengaged. It takes quite a bit for make a system real time instead of non-real time. The front end would need to send out a request to the and update the results based on the server. This would need to continuously happen every second while the program is active. The Database likewise would need to update in real time to reflect the students engagement status to be then reported. This would cause difficulties in the system because of the processing speed this would require and in many places the system could have errors.

2.3. Finding Effective Engagement Detection Algorithm

Our final challenge was to find an effective engagement detection algorithm [13]. We went through a lot of papers. Some papers appear to be good yet have no effective evaluation methods. Others did not have their own training data sets. Since we lacked our own training data we could not sufficiently train our own algorithms. A combination of these factors and more made choosing detection algorithms difficult.

3. SOLUTION

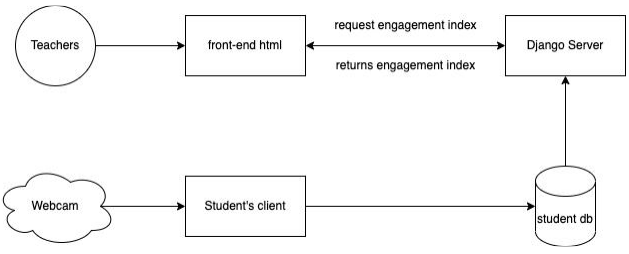


Figure 1. Overview of the solution

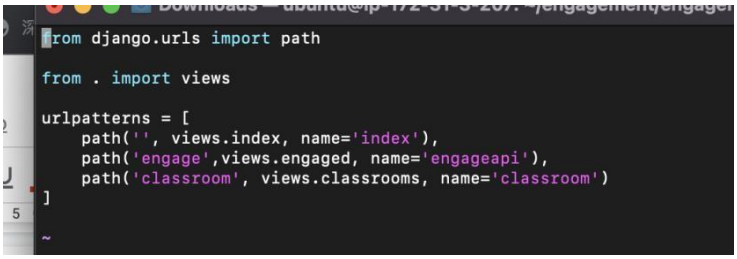
Our solution is a web based interface based on a student side run algorithm. The process begins with multiple users, 1 teacher, and 1 or more students. First the student side client uses the webcam to collect the raw data. This data is then passed through the engagement algorithm which processes the data into student engagement. The student engagement is then sent to a student database where it will be stored for this session. The student database then gets stored onto a Django Server where it can be more readily accessible. The teachers side of the solution is a web page where each student is listed by name with an index next to it to demonstrate engagement. To better assist the teachers in recognizing when a student disengages, The font and color of the row the student's data resides on will change to better visually alert the teacher. This front end web site is in constant communication with the Django server requesting engagement indexes. The Django server which is constantly being updated with new data from the student database will then provide the front end by returning the requested indexes. This new data is then shown on the web page which then alerts the educator to any potential disengagement from the students.

```
                gaze_weights = 2

# Concentration index is a percentage : max weights product = 4.5
        concentration_index = (
            emotionweights[self.emotion] * gaze_weights) / 4.5
        print("this is raw ci: ",concentration_index )

        self.conn.sendStudentEngagementInfo("Eric", concentration_index)
        if concentration_index > 0.65:
            return "You are highly engaged!"
        elif concentration_index > 0.25 and concentration_index <= 0.65:
            return "You are engaged."
        else:
            return "Pay attention!"
```

Figure 2. Upload the weight to the server

```
from django.shortcuts import render
from sever.models import student, raw_reading, event_table, classroom
# Create your views here.
from django.http import HttpResponse


def index(request):
    stu = student.objects.all()
    print(stu[0].name)
    cla = classroom.objects.all()
    return render(request, 'class.html',{'students':stu, 'classrooms':cla })



def engaged(request):
    stu = request.GET['stu']
    print("student name from sever: ",stu)
    student_1 = student.objects.filter(id =stu)
    print("studnet status:"+ str(student_1[0].engagement))
    return HttpResponse(student_1[0].engagement)

def classrooms(request):
    classes = request.GET.get('classroom')
    stu = student.objects.filter(classroom__name = classes)
    cla = classroom.objects.all()
    return render(request, 'student.html',{'students':stu, 'classrooms':cla})
#   return HttpResponse("Hello,"+ stu[0].name )
~
~
```

Figure 3. Screenshot of code 2



```
from django.urls import path

from . import views

urlpatterns = [
    path('', views.index, name='index'),
    path('engage',views.engaged, name='engageapi'),
    path('classroom', views.classrooms, name='classroom')
]
```

Figure 4. Screenshot of code 3

```
"""
from django.contrib import admin
from django.urls import path ,include

urlpatterns = [
    path('admin/', admin.site.urls),
    path('' , include('sever.urls')),
]
```

Figure 5. Screenshot of code 4

```
        seconds = 1
      var stu{{student.id}};
        stu{{student.id}} = "{{ student.id }}";
        setInterval(function () {
      $.ajax(
      {
          type:"GET",
              url: "http://13.57.184.27:8000/engage",
          data:{
                  stu: stu{{student.id}}
          },
          success: function( data )
          {
              console.log(data);
              if (data == "False") {
                  $(".stu{{student.id}}").css('color', 'red');
              }
              else {
                  $(".stu{{student.id}}").css('color', 'black');
              }

              $(".stu{{student.id}}").text(data.toString());
              //$( '#message' ).text(data);
          }
          })
       }, seconds * 1000)
```

Figure 6. Screenshot of code 5

4. EXPERIMENT

4.1. Experiment 1

This experiment would be completed during a small group class where all students are on the engagement algorithm. The class would be completed as usual and the student's average engagement will be gathered by setting engaged as 1 and disengaged as 0 and finding the average for each student. After the session is completed, Students will be asked for how engaged they were during the class on a scale of one to ten. These two percentages will be used to find the percent error of the detected engagement value to determine the accuracy of the algorithm.

Class 1	Student 1	Student 2	Student 3
Detected result	74.1	38.2	86.4
claimed result	80	60	90
%Error	-7.375	-36.33333333	-4

Figure 7. Table 1

Class 2	Student 1	Student 2
Detected result	68.3	58.9
claimed result	60	70
%Error	13.83333333	-15.85714286

Figure 8. Table 2

As the Data shows, The algorithm used is mostly effective at detecting the engagement of the students. There are some outliers in the data, but due to the previously mentioned diversity in Neurodivergent engagement, this was to be expected.

4.2. Experiment 2

This experiment will be accomplished in the form of a survey after a few classes with the system. Teachers will be asked to answer a few questions about their experience with the tool, and answers will be recorded on a scale of Strongly disagree to Strongly agree. This survey will include Questions about the system effectiveness and usability. We will be using the System usability scale.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I think that I would like to use this system frequently.	X				
I found the system unnecessarily complex.				X	
I thought the system was easy to use.			X		
I think that I would need the support of a technical person to be able to use this system.		X			
I found the various functions in this system were well integrated.		X			
I thought there was too much inconsistency in this system.				X	
I would imagine that most people would learn to use this system very quickly.		X			
I found the system very cumbersome to use.					X
I felt very confident using the system.		X			
I needed to learn a lot of things before I could get going with this system.				X	

Figure 9. Survey

In the above two experiment we have proved that our tool accomplished it intended work by being accessible to teachers and decently accurate for the students.

5. RELATED WORK

This work is a good work on the effects of autism spectrum disorder [4]. The paper generally talks about the experiences and effects living with ASD while our paper is specifically on a possible way to better aid the education of neurodivergent children. Our paper does not specifically touch on the actions implications of ASD, we do give a generalization of ASD, especially the effects it has on a child's education and their ability to function in a learning environment.

This is the paper which contributed the algorithm of the paper [3]. The related paper was entirely on creating an algorithm which can detect a students engagement in a classroom setting. This is done through a combination of eye gaze and facial expressions. Our paper is less about the

detection algorithm and more about the web based interface and how it benefits neurodivergent students. Their work is better for understanding engagement detection and how it works, while ours is more focused on the application in education.

This work is a good work on the correlation of engagement and education [5]. This work mainly focused on the connection between a student's engagement in a class, and their ability to retain the information they learned about. Our paper on the contrary is about how to get the students to be more engaged in a classroom setting. Both papers deal with the concept of education and by nature places education in a place of high importance.

6. CONCLUSIONS

In Summary we have created a Tool which is specifically designed to assist in the education of the neurodivergent youth. This program uses A simplistic Engagement detection algorithm to provide a general estimate of student engagement and provides this engagement data to the teacher in real time to assist the teachers in keeping students as engaged as possible to help them be as efficient as possible in the learning process [14]. We then put this tool to the test in two different experiments to demonstrate the accuracy of the system on the students end and the ease of usability on the teachers end. Both of these result proved to be satisfactory and showed that our tool is effective and not difficult to maneuver.

Although our Solution is effective in most scenarios, it is not without its flaws and issues. One of these issues is that due to the natures of the attentiveness within the community of the neurodivergent youth, this program could not possibly work for all cases.

Due to this limitation we decided to choose a very simplistic algorithm as it could apply to more people and could be less affected by the individual quirks of the students' engagement or disengagement [15]. Due to this simplistic algorithm, the programs in certain cases could really return false positives and negatives with students attentiveness. Additionally, The practicality of the system can still be wildly improved as it currently takes downloading the software onto the student devices meaningthat setup time can be the length and riddle with bugs.

Most of these limitations can be solved with more time and resources. Some solutions can be currently investigated include creating a more accessible student user interface to make setup easier, creating delineations to allow multiple classes to be ran at the same time, or even to move the entire system online so no downloads are required making new student initiation much simpler and faster.

REFERENCES

[1] Bryson, Susan E., Sally J. Rogers, and Eric Fombonne. "Autism spectrum disorders: early detection, intervention, education, and psychopharmacological management." The Canadian Journal of Psychiatry 48.8 (2003): 506-516.
[2] Harrower, Joshua K., and Glen Dunlap. "Including children with autism in general education classrooms: A review of effective strategies." Behavior modification 25.5 (2001): 762-784.
[3] Sharma, Prabin, et al. "Student engagement detection using emotion analysis, eye tracking and head movement with machine learning." arXiv preprint arXiv:1909.12913 (2019).
[4] Lord, Catherine, et al. "Autism spectrum disorder." The lancet 392.10146 (2018): 508-520.
[5] Jablon, Judy R., and Michael Wilkinson. "Using engagement strategies to facilitate children's learning and success." YC Young Children 61.2 (2006): 12.
[6] Spiel, Katta, et al. "Adhd and technology research–investigated by neurodivergent readers." Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022.
[7] Spiel, Katta, and Kathrin Gerling. "The purpose of play: How HCI games research fails

neurodivergent populations." ACM Transactions on Computer-Human Interaction (TOCHI) 28.2 (2021): 1-40.
[8] Sibert, Linda E., and Robert JK Jacob. "Evaluation of eye gaze interaction." Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 2000.
[9] Spiel, Katta, et al. "Adhd and technology research–investigated by neurodivergent readers." Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022.
[10] Sasson, Noah J., et al. "Neurotypical peers are less willing to interact with those with autism based on thin slice judgments." Scientific reports 7.1 (2017): 1-10.
[11] Biederman, Joseph. "Attention-deficit/hyperactivity disorder: a selective overview." Biological psychiatry 57.11 (2005): 1215-1220.
[12] Moore, Joi L., Camille Dickson-Deane, and Krista Galyen. "e-Learning, online learning, and distance learning environments: Are they the same?." The Internet and higher education 14.2 (2011): 129-135.
[13] Kramer, Thomas R. "Pocket milling with tool engagement detection." Journal of Manufacturing Systems 11.2 (1992): 114-123.
[14] Kramer, Thomas R. "Pocket milling with tool engagement detection." Journal of Manufacturing Systems 11.2 (1992): 114-123.
[15] Kahn, William A. "Psychological conditions of personal engagement and disengagement at work." Academy of management journal 33.4 (1990): 692-724.

A SMART PLANT MOISTURE LEVEL DETERMINATION SYSTEM TO DETERMINE IF THE PLANT NEEDS TO BE WATERED OR NOT BY USING MACHINE LEARNING

Ruohan Zhang[1], Yaotian Zhang[1], Yu Sun[2]

[1]Fairmont Preparatory Academy, 2200 W Sequoia Ave, Anaheim, CA 92801
[2]California State Polytechnic University, Pomona, CA, 91768, Irvine, CA 92620

ABSTRACT

All living things including plants need water to survive, and agriculture is the world's biggest user of water [4]. Unfortunately, in a worst-case scenario, over-watering and drying up cause both water waste and the plant's death [5]. Guided by this problem that is frequently occurring around the world, we designed an app to determine if the plant needs to be watered or not by capturing pictures of a certain plant and training an AI to compare whether the soil in the pot is dry or wet. In this program, we use Raspberry Pi to capture an image of the plant every 10 seconds,in which the Python code using TensorFlow inside the Raspberry Pi will determine the moisture level of the soil [6][7]. The result will be posted to Firebase with a timestamp, and lastly, we have a mobile app that can display the result from Firebase to the user. We published our application to Apple's App Store and the Google Play Store, and public installation of the app means that it can have more widespread usage. An experiment was performed to determine whether the application's model can accurately determine soil moisture [8]. The results indicate that the model is very accurate for the vast majority of soil samples under various lighting conditions.

KEYWORDS

Python, Flutter, Machine Learning, Firebase

1. INTRODUCTION

If you ever noticed your plant is turning yellow, it is possible that your plant is being overwatered. Overwatering is one of the most common causes of plant problems [9]. Overwatering severely limits the supply of oxygen that roots depend on to function properly, meaning that plants do not get enough oxygen to survive. Also, if the soil is heavily drained, it will become waterlogged and the roots growing in this soil will die [10]. Furthermore, overwatering can lead to broader problems such as the over usage of water. Take California farmers as an example. In the year 2021, California's farmers pumped an additional six to seven million acre-feet of water from their wells above what they normally use. This quantity of water would cover 10,000 square miles with a foot of water. Problems related to overwatering are happening in a lot of parts of the world, from one's backyard garden to a local farm. So, in what ways can people prevent a plant from being overwatered or dried out?

Some people come up with a sensor that can detect the moisture level of soil for the purpose of avoiding overwatering. However, these sensors assume that the owner only has a small number

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 127-134, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130410

of plants, which is not the case. Normally, people with large farms tend to need these sensors more because they can't take care of every plant that well and make sure they are healthy. To make sure every part of the soil on the farm is healthy, they will need to buy hundreds and thousands of them, which leads to the second issue. A second practical problem is that the sensor needs to be taken care of. These sensors detect the moisture level by directly inserting them into the soil and waiting for a few minutes then pulling them out. Let's use the case as if you own a large plantation and are rich enough to afford to buy many of these sensors. You will need to insert them one by one and sit there waiting for the sensor to work. Then, you will need to record the mixture level for each area you measure and determine whether a certain area needs to be watered or not. Next, you will go around your plantation again to pull each sensor out of the soil, and lastly, you need to clean them for them to work again. This process takes both time and effort, which are both valuable. Seeing these issues with existing tools, we came up with our topic of creating an app that can monitor your plant and notify you when the plant needs to be watered, which can avoid both overwatering and drying out of the plant.

The purpose of our application is to predict and provide a real-time moisture level of plants and avoid overwatering or withering. To provide an accurate estimate, the application uses many steps to make predictions. Firstly, the application gathers the plant's picture by using a small Raspberry Pi camera that the user can operate simply. Secondly, the Raspberry Pi camera sends the picture to Raspberry Pi, where our program analyzes and processes the image. Thirdly, Raspberry Pi sends the results and analyzed data to Firebase, where our server is built and data is stored in [11]. Lastly, Firebase returns the results to the user's mobile application where the user can access the data. However, only following a specific order to process the image in some cases won't always be accurate, instead, we also used a machine learning process to improve the application even more. Compared to other moisture sensors, our sensor requires much fewer conditions to run accurately, for example, our sensor can take pictures of large amounts of soil by simply pointing the camera to it. The sensor analyzes the soil as soon as the user takes the picture. While some other moisture detectors requires other more complicated steps like pointing a long iron stick, our method has a much simpler and more streamlined process.

The effectiveness of the application can be measured by the accuracy of the application in determining the soil moisture of a given sample of soil. Implementing a smaller-scale experiment to start with can pave the way to future larger-scale experiments after the necessary adjustments to the application have been made. The experiment involves 20 different soil samples, of which 10 of them are dry and 10 of them are thoroughly watered. Using the application, each sample will be analyzed for its soil moisture by taking a picture of it from a top-down angle. Having all the pictures taken from the same angle can reduce confounding variables in the experiment. The total number of dry soil samples and the number of wet soil samples that were correctly identified will be recorded in a table. The goal of this experiment is to ensure that the basic code and model within the application work as intended so that more adjustments and expansions can safely be made to the application in the future. If almost all of the soil samples are identified correctly by the application, the code and model will need no further adjustments in the near future, and effort spent on improving the application can be applied to other aspects of the application instead. However, if a significant portion of the samples is incorrectly identified by the application, then the most urgent change to make would be creating a new model or adjusting the current model. By progressing with the application, it can hopefully find practical use in the field of agriculture.

The rest of the paper is organized as follows: Section 2 gives the detail on the challenges that we met during designing and developing the application and the experiment to test the effectiveness of the application; Section 3 focuses on the detail of our solution related to the challenges that are mentioned in section 2; Section 4 presents the details about the experiments we did and the

related works will be presented in Section 5; lastly, Section 6 provides concluding thoughts regarding the project as well as a brief self-reflection to see what could be improved on in the near future.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Deciding what angle to approach the problem with

The first obstacle with the project was deciding what angle to approach the problem with. Providing plants with an improper amount of water to cause overwatering and drying is a relevant issue, so a reasonable solution would be offering people a method to measure the moisture of a given area of soil. To do this, a sensor could be used to take a picture of the soil. Using the picture of the soil, a system would need to determine the soil moisture from the picture and return the result to the user in a convenient manner. The simplest solution to do so is a mobile application. Most people carry a phone around nowadays, which means that almost everyone will be able to easily access the application. The main overall concept behind the mobile application is retrieving the soil moisture from the sensor and the back-end code, then printing the predicted soil moisture to the screen, which can inform users of the application and help them determine whether or not they should keep watering a plant.

2.2. Creating the code that would be used in the application

The next challenge is creating the code that would be used in the application. The purpose of the code would be to retrieve a picture of the soil and make a prediction as to whether the plant has been properly watered or not. To do so, a sensor that acts as a camera will constantly check what it is currently seeing for soil moisture by using a while True loop. By using a while True loop, the code ensures that as long as the application is running, the application will constantly run the lines of code to update the image every 10 seconds with the most recent image that the sensor detects, then loads a new model to run through the updated image with. Another issue with the code is that the front-end is made with Thunkable and the back-end is coded using Flutter. To combine the two different programming languages, a Flask server is used; the Flask server helps HTTP requests move back and forth. Furthermore, Firebase also helps with the transferring of image files.

2.3. Figuring out how to experiment with the system

The final obstacle was figuring out how to experiment with the system. The ideal experiment would be testing with multiple sensors across a wide area of land for soil moisture. However, the application is currently only capable of supporting one sensor at a time, and access to much farmland is costly and difficult to acquire. Furthermore, before making such a large-scale experiment, determining whether the application is reliable at accurately gauging soil moisture on a smaller scale is an important step. Therefore, the experiment that was decided upon was to take ten samples of dry soil and ten samples of moisturized soil, use the sensor to observe the predicted soil moisture levels from the application for each sample, and record the results on a table. If the application was able to accurately identify the soil moisture for the vast majority of the soil samples, it could be concluded that the application would make for a reliable product.

3. SOLUTION

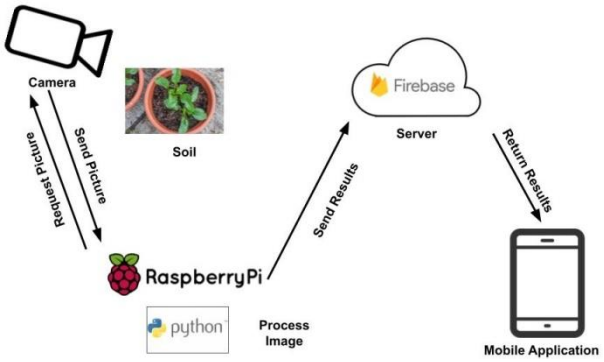


Figure 1. Overview of the solution

The system of the application involves a camera, a Raspberry Pi, a Firebase server, and the mobile application itself. A Raspberry Pi is a small and low-cost single-board computer that is capable of requesting and retrieving pictures from a connected camera. The camera is placed to face a soil sample, and the camera takes a picture of the soil and sends it to the Raspberry Pi whenever it is instructed to do so. The Raspberry Pi runs Python code that acts as the back-end of the mobile application and runs a model to determine whether the soil sample that is captured in the picture is dry soil or wet soil, using the image retrieved from the camera as input for the model [12]. As Python is a popular and relatively simple language in terms of syntax, there are multiple packages to choose from dedicated to artificial intelligence and image classification. After the model finishes processing the image, the results are sent to the Firebase server. Finally, the results are retrieved from the server and printed in the mobile application. The front-end of the application is created with Thunkable, which is a platform that prides itself on its simplicity and allows the building of mobile applications with little to no code; rather, Thunkable users can drag and drop any components of the user interface that they desire. Within the application, the information that is projected to the screen is the classification of either wet soil or dry soil, the percentage of confidence from the model, and the image that was taken from the camera.

Python was selected to be the code for the back-end; it is not only easy to work with due to its relatively simple syntax but also convenient to achieve specific features due to the myriad of packages that are available. To create the back-end of the application, the Python code was separated into four separate Python files. The first Python file is for the application itself. Within this file, the cv2 file is used to perform video capture [14]. Then, a while True loop is made to indicate that the code within this loop should continue to run as long as the application is active and running. Variables collect what is currently being read from the camera, and the timestamp is collected as well by using Python's time package. Every 10 seconds, the old timestamp will be replaced with the new, current timestamp, and a new image will be retrieved. Using the path of the new image, the predict method from the classifier file is called. The predict method makes use of three other methods; one of them loads a model, the second one loads the labels, and the final one loads the image. Using the package TensorFlow, a model is loaded in and allocated tensors [13]. Then, the input and output tensors are retrieved and returned within the method to load a tflite model. For the method that loads labels, the path to the labels is opened and read, then compiled into a list and returned. The final method to load the image uses an image path and specifies that the target size of the image should be 224 pixels by 224 pixels. The loaded image is then converted to an array, has a batch created for it, and then is returned. The predict method loads the model, loads the label, and loads the image in that order. After the input tensor is set, the inference is run by calling the invoke method on the model, then the prediction is

retrieved by retrieving the tensor and taking the one with the highest confidence.

Thunkable is used as the front-end, and the application is divided into two separate screens. The first screen simply has a logo object that acts as an introductory splash screen. The second screen is the main screen that reveals all of the needed information to the user. There are four components on this screen, which are the soil label, the time last updated in seconds, the confidence label, and the captured image by the camera. The soil label states whether the soil sample was determined to be a “Wet Soil Pot” or a “Dry Soil Pot”. The confidence label states how confident the model was in its prediction as a percentage. The Firebase real-time database is indicated within Thunkable as an invisible component, which is how the components on the second screen are able to retrieve the necessary information.

```python
@app.route('/')
def index():
    #return render_template('index.html', filename='uploads/endgame_trailer_final.mp4')
    return 'Soil Image Classification'

@app.route('/classify_image', methods=['GET', 'POST'])
def classify_image():
    if request.method == 'POST':
        f = request.files['image']
        print(f)
        filename = 'images/'+ f.filename
        print(filename)
        f.save(filename)
        pred = classifier.predict(filename)
        if pred != None:
            os.remove(filename)
            return json.dumps(pred)

        return 'Image cannot be analized'
```

```python
def upload_file(self, storage_filename, local_path):
    bucket = storage.bucket()
    blob = bucket.blob(storage_filename)

    if blob.exists():
        print('This file already exists on cloud.')
        return blob.public_url
    else:
        outfile = local_path
        blob.upload_from_filename(outfile)
        with open(outfile, 'rb') as fp:
            blob.upload_from_file(fp)
        print('This file is uploaded to cloud.')
        blob.make_public()
        return blob.public_url
```

Figure 2. Pictures of the Python back-end code

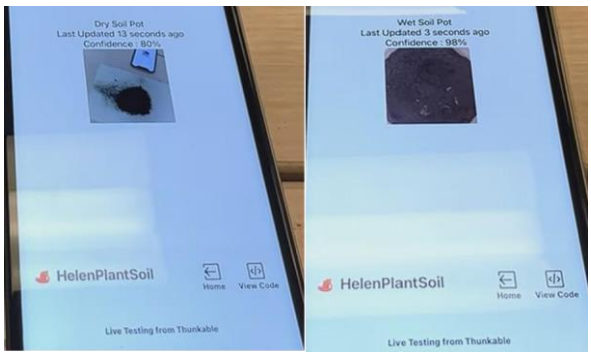


Figure 3. Pictures of the mobile application

4. EXPERIMENT

An experiment was performed specifically to test the accuracy of the model in correctly identifying the moisture of soil in a given soil sample. Ten wet soil samples and ten dry soil samples were gathered, making twenty different soil samples in total. To reduce confounding variables regarding the samples themselves, each sample used approximately the sample amount of soil, and each wet soil sample was watered with approximately the same amount of water. Then, the sensor in the application system was used to take pictures of the soil from a top-down angle, in which taking the pictures of all the samples from the same angle further reduces confounding variables by keeping each sample as consistent as possible. After all the samples have been tested for their soil moisture, the number of correctly identified samples is recorded in a table.

Soil Moisture Type	Number of Correctly Identified Samples	Number of Total Samples	Accuracy
Wet	10	10	100%
Dry	9	10	90%

Figure 4. Table of experiment 1

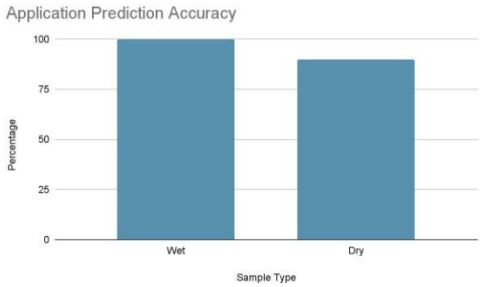


Figure 5. Application prediction accuracy

The application scored highly on its accuracy of both dry soil samples and wet soil samples. The experiment indicated that the application has a 100% success rate when it came to identifying wet soil samples. However, the application only scored a 90% success rate when testing dry soil samples, in which one of the soil samples was identified incorrectly as a wet soil sample. A possible explanation as to why one of the dry samples was identified incorrectly was because of inconsistent lighting throughout the duration of the experiment. The samples were tested outdoors, and the sun changes position and casts different amounts of light on the soil samples throughout the day. From the samples that were tested, the wet samples were noticeably darker than the dry samples. Because consistent lighting was not managed in the experiment, inaccuracies may be caused by the model not being well-trained enough at identifying samples under different levels of lighting.

The experiment was designed to test the model within the application for its accuracy in identifying the moisture in samples of soil, and the results of the experiment would indicate whether the model needs improvement and whether more efforts should be focused on improving the model or improving other aspects of the application. According to the results of the experiment, the model appears to do well at determining the moisture of both wet and dry soil samples. This falls within expectations, as the model has already undergone much training with various soil samples and was expected to perform fairly well. As previously mentioned, something that was intentionally left as a possible confounding variable is the lighting in the

pictures of each sample. In a real-life agricultural setting, the lighting of the soil when detected by sensors would not always stay the same, and the application would need to be tested for its ability to handle different levels of lighting as a result.

5. RELATED WORK

A related work describes different methods that are used to evaluate the moisture of soil, which vary depending on the application or setting that it is used. Therefore, a concept was proposed so that different approaches could be combined into a single integrated system that could be potentially used as a multipurpose solution [1]. This work is similar to the related work in that the primary focal point is analyzing the moisture of the soil. However, the related work goes into depth on various methods that could be used to quantify the moisture of the soil. On the other hand, this work emphasizes the creation of an application to easily gauge whether the soil is properly watered or not.

In another related work, different types of sensors are compared regarding their ability to accurately measure soil moisture. Between the TDR-based sensor that made use of the travel time of an electromagnetic pulse to propagate along sensor rods and the 10HS sensor that worked through capacitance, it was concluded that both types of sensors had their shortcomings and neither one definitively outperformed the other [2]. The related work and this work share the similarity of determining soil moisture. However, while the related work compares how effective different sensors are at evaluating soil moisture, this work focuses on incorporating a sensor into an application.

A third related work experiments on the application of using soil moisture in a drip irrigation automation system that was primarily composed of a base station unit, a valve unit, and a sensor unit. The system was tested on an 8-decare area with dwarf cherry trees, and it was observed that the system was low-cost and reliable and could have practical agricultural use [3]. Both this work and the related work were similar in that the goal was to create an application using sensors to detect soil moisture that would hopefully have practical use in agriculture. However, the related work involves performing a large scale experiment on a wide area of land while this work aims to test the accuracy of the sensors instead.

6. CONCLUSIONS

The method that has been implemented to resolve the issue of improperly watering plants is a mobile application that can tell its users whether a sample of soil is wet or dry. Recognizing that the soil is dry encourages the users to water their plants, and recognizing that the soil is wet can inform the users that there is no need to water their plants for the time being. By using this application, people can prevent overwatering the plants and accidentally killing them; on the other hand, they can also potentially be alerted to the fact that the plants may not be getting enough water. The application was tested in an experiment in which ten dry soil samples and ten wet soil samples were used to take pictures for the application. The samples were taken outside at various times during the day, which ensured that the soil sample pictures were taken at different lighting levels. The number of times that the application correctly determines the moisture of the soil was recorded in a table separately based on whether the tested soil sample was a wet or dry sample. According to the results, the application's model is very proficient at determining the soil moisture. Because the lighting levels were different across each picture, the model has proven to be somewhat robust across multiple lighting conditions. However, the single inaccuracy of the model from the experiment indicates that while it is not an urgent issue, the model in the application still has room for improvement.

One of the most significant current limiting factors in the application is its ability to use multiple sensors. Currently, only one sensor at a time can be used with the application. However, in a more realistic agricultural setting in which more farmland would have to be analyzed for its soil moisture, the application may be impractical to use [15]. To keep the application relevant within the agricultural field, more time and effort would need to be spent to allow the application to take in multiple sensors at a time and do so in a manner that still keeps the user interface clean and easy to navigate.

Something that could be done is adjusting the sensor page to contain information from multiple sensors. As the current sensor page shows what is being seen by the sensor, the page wouldlikely have to be scrollable so that the user will be able to see a live feed from multiple sensors in one place.

REFERENCES

[1] Schmugge, T. J., Jackson, T. J., & McKim, H. L. (1980). Survey of methods for soil moisture determination. Water Resources Research, 16(6), 961-979.
[2] Mittelbach, H., Lehner, I., & Seneviratne, S. I. (2012). Comparison of four soil moisture sensor types underfield conditions in Switzerland. Journal of Hydrology, 430, 39-49.
[3] Dursun, M., & Ozden, S. (2011). A wireless application of drip irrigation automation supported by soil moisture sensors. Scientific Research and Essays, 6(7), 1573-1582.
[4] Lichtenberg, Erik. "Agriculture and the environment." Handbook of agricultural economics 2 (2002): 1249-1313.
[5] Koop, Steven HA, and Cornelis Johannes van Leeuwen. "The challenges of water, waste and climate change in cities." Environment, development and sustainability 19.2 (2017): 385-418.
[6] Zhao, Cheah Wai, Jayanand Jegatheesan, and Son Chee Loon. "Exploring iot application using raspberry pi." International Journal of Computer Networks and Applications 2.1 (2015): 27-34.
[7] Abadi, Martín. "TensorFlow: learning functions at scale." Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming. 2016.
[8] Schmugge, T. J., T. J. Jackson, and H. L. McKim. "Survey of methods for soil moisture determination." Water Resources Research 16.6 (1980): 961-979.
[9] Morton, T. G., A. J. Gold, and W. M. Sullivan. Influence of overwatering and fertilization on nitrogen losses from home lawns. Vol. 17. No. 1. American Society of Agronomy, Crop Science Society of America, andSoil Science Society of America, 1988.
[10] Bradford, Kent J., and Theodore C. Hsiao. "Stomatal behavior and water relations of waterlogged tomato plants." Plant Physiology 70.5 (1982): 1508-1513.
[11] Moroney, Laurence. "The firebase realtime database." The Definitive Guide to Firebase. Apress, Berkeley, CA, 2017. 51-71.
[12] Islam, Rashedul, Rofiqul Islam, and Tohidul Mazumder. "Mobile application and its global impact." International Journal of Engineering & Technology 10.6 (2010): 72-78.
[13] Pang, Bo, Erik Nijkamp, and Ying Nian Wu. "Deep learning with tensorflow: A review." Journal of Educational and Behavioral Statistics 45.2 (2020): 227-248.
[14] Weiss, Patrice L., et al. "Video capture virtual reality as a flexible and effective rehabilitation tool." Journalof neuroengineering and rehabilitation 1.1 (2004): 1-12.
[15] Ochsner, Tyson E., et al. "State of the art in large-scale soil moisture monitoring." Soil Science Society ofAmerica Journal 77.6 (2013): 1888-1919.

LEXICAL FEATURES OF MEDICINE PRODUCT WARNINGS IN THE PHILIPPINES

Shielanie Soriano-Dacumos

University of Rizal System, Binangonan, Rizal Philippines

ABSTRACT

Over the stretch of years, the Philippines has been facing numerous medical problems since the public outcry against a 'dengue' vaccine. As a result, parents refused their children from having an anti-measles vaccine which created a medical outbreak in the country. Product warnings are found to be in their optimal position in safeguarding the life of consumer-patients. This paper anatomizes the lexical features of medicine product warnings in the Philippines which are crucial in the response discourses. A range of linguistic frameworks were applied and significant findings were drawn. Lapses on the use of noun abstractness, synthetic personalization, field continuum, adjectives, and adverbs were identified. Such an investigation brought up the transparency of communicative features of medicine safety texts. In the end, linguistic components create a vital impact on the legal content adequacy of medicine product warnings, unfolding the vitalities of these messages in facilitating informed decision-making among consumer-patients.

KEYWORDS

Medicines, Consumer-patients, Linguistic components, Product warnings

1. INTRODUCTION

Consensus exists in the body of research agreeing that product warnings must contain four elements (Heaps & Henley, 1999; Sanders & McCormick, 1993): a signal word, a hazard statement, a statement of consequences, and instructions for avoiding the hazard. If product warnings are badly planned, Tiersma (2002) accentuated that product manufacturers are held liable to law. Legal responsibilities for damages strengthen the case of manufacturer's negligence on his failure to provide effective product warnings among consumers.

Product warning texts are fully applicable to the civil context and to every detail as interesting as any criminal case. In the case of the Philippines, product recall is the immediate legal solution once the hazard has been known to Food and Drug Authority (FDA) or Department of Trade and Industry (DTI). These government agencies instantly warn the public through mass media about the discovered chemical risks in the product. In line with this, the FDA in its official website, identified essential standards for setting warning information. Among these are: 1.) statements must be prominent and conspicuous; 2.) the FD&C Act and related regulations specify warning and caution statements related to specific products; 3.) cosmetics that may be hazardous to consumers must bear appropriate label warnings (e.g. flammable cosmetics).

However, there are some cases that product warning readers do not follow the instructions stated on the label because they do not understand the written directives. The study of Horowitz (1985) disclosed that great consideration should be given both to skilled and unskilled consumers. The warnings should alert them to read and understand the safety instructions prior to the utilization

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 135-149, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130411

of the product thus allowing them to make informed decisioin by considering the readability (Didonet and Mengue, 2004) and comprehensibility (Hancock, Rogers, et al., 2004) of the warning texts. Product warning researchers (Wogalter, 2006; Malik, 2002; Shuy, 2000) further attended to the issues of product users having lower language and communicative skills. Language differences in product warnings are also to be considered in the warning literature. Arai (2002) on the other hand, implies the essence of bilingual warnings to be applied in medicines since it will ensure the compliance of linguistic minorities, and may result to safety submission.

1.1. Medicine Product Warnings in the Philippines

According to Go (2001), medicines, being regulated goods, cannot be treated as mere trade commodities but should be managed as a health utility. Akin to his statement, Philippines was bannered in various newspapers' headlines globally in April 2016 as the country's Department of Health (DOH) vaccinated 830,000 elementary schoolchildren with the first ever vaccine which claimed to fight dengue (Philippine Daily Inquirer, 2018), a possible life-threatening mosquito-borne disease that can cause an unsafe drop in blood pressure and life-threatening bleeding. After a year, such vaccination initiative by the government brought panic among parents as it is labelled by Philippine Safety Advocates (2017) in their media campaign that such program is the biggest government funded clinical-trial-masked-as-a-public-health-program scam of an experimental drug in the history of the DOH. From here, the Philippines' Food and Drug Authority (FDA) withdrew the approval of the vaccine.

With the government's concern to safety, the question of not determining the product risks prior to vaccine's distribution became the major concern of the panicking public specifically the parents whose children had been vaccinated by such dengue preventive. Moreover, the issue of not incorporating the hazards on its product warnings created a major dismayed to the warning researchers since the government failed to safeguard the public's right to know and the drug authorities' liability for consumer safety.

In November 2013, Philippine Star featured an article highlighting the FDA's order to recall four products of pharmaceutical firm Eli Lilly (Philippines) Inc. after they are found to be unregistered. Most of their products are anti-depressants.

Another crucial issue on medicine risks was also posted on the DTI's website which discloses the importance of honest information in the dietary food supplement which could be bought in grocery stores. The caption "No Approved Therapeutic Claim" is suggested to be printed on the main display panel of all labelling materials used in food supplements (i.e. the immediate label of the bottle, drug box, carton, information sheet, leaflets, etc.) to guarantee that these consumer products are not commercially sold or publicized with some therapeutic claims.

Alarmingly, the volume of safety information presented are crucial in the analyses of the real-world effectiveness of product warnings. Emphasizing that the clarity of language is essential in providing product warnings, RA 7394, Art. 7 of Philippine Constitution generally promulgates the inclusion of product caution having clear and adequate safety warnings or instruction, or requirements. However, product liability attorneys and government agencies oftentimes concentrate on product defect cases, and not on the presumed negligence of the product warning writers or the product manufacturers. The drug safety information specifically the product warnings served as the most immediate reference of parents and health personnel in identifying the risks associated in medicine use. Manufacturers' conformity on the safety components of warning texts is a major call for government officials in the Philippines to closely monitor and strictly implement.

1.2. Forensic Linguistics

The demarcation line which separates business and law is obliterated when product liability cases started gaining ground. Distinctively, the use of language is carving a place in the field of law and has resulted in the framing of Forensic Linguistics which is the crossing point between language and law. Olsson (2006) provided proof that the law is bounded with the police force, court trial matters, legislation, product or property disputes, court proceedings, or some inevitable real life situations which look for a legal remedy. Language, in most cases, is extracted and thereby becomes evidence in court.

In 1993, Shuy identified various crimes associated with language, which includes physically nonviolent crimes of bribery, solicitation to murder, sex solicitation, business fraud, selling or purchasing stolen property, perjury, threatening, and importantly product warnings. Written or spoken language, both entail textual tenets which serve as strong evidence in court.

Linguistic terrains in the domains of law continue to grow. Moenssens, et. al.(2000) name forensic linguistics as an evaluation of the linguistic characteristics of communication. Since language is an effective tool in understanding the complexities of legal language and interpretation, then grammar, syntax, spelling, vocabulary, and phraseology are helpful evidence in court. Tiersma (2002) emphasizes that forensic linguistics is the usage of linguistic acumen and methodologies in solving factual issues that are relevant to legal disputes. From here, a great deal of interest in the intersection line between language and law is drawn, particularly in the product warning discourse.

Evidently, language complexities enable the interconnection of the two gigantic genres of the marketplace--law and business. Evaluating the word, spelling, grammar, phrases, syntax and even punctuation marks is a common analytical aspect of language and these are all evident in the writing of product warnings which is basically under the umbrella of product liability.

1.3. Language and Communication

In communication, two or even more individuals send and receive messages and could be encoded automatically by human language. The lexical features of communicative health information are made available to consumers and therefore analyzed to determine the safety level of product warnings.

Safety communication could be achieved if the message is evidently clear and will totally warn the product users about the imminent hazards. The manufacturers as senders of the message are held liable in the construction of the warning texts. The message must be available and comprehensible to the consumers as target receivers of communicative texts. If the transmitted message is inadequate, receivers can formulate limited information about the products, which in the end, an erroneous impression can be perceived.

Since the goal of language is to establish a good sense of communication between the sender and receiver, Shuy (2005) highlights that communication necessitates a specified quantity of assigned data in order to impart the framework essential for product consumers to comprehend what is being shared with them. On the surface, it is with words that information is apprehended. When the message of the sender creates a misleading impression to target receivers, it leads to wrong inferences and conclusions, and ultimately it results to communication breakdown.

1.4. Research Question

Language plays a pivotal role in transmitting health-communication information, likewise it determines the adequacy issues of existing product warning texts available in the mainstream supermarkets and drugstores. This research posits the question: What are the lexical features of selected medicine product warnings in the Philippines? Such an inquiry brings to the fore some significant implications for the adequacy requirements of product warnings, realizing the vitality of cautionary texts for consumers.

1.5. Theoretical Framework

To bring out the optimization of the linguistic features of the warning texts, the study investigated the lexical attributes of medicine product warnings.

In analyzing the lexical features, the usage of Lyon's noun entities (1977) was applied. Moreover, Halliday's Words in Field Continuum (1993) plays a vital portion in categorizing the classification of words. In the use of signal words, this study mirrored what Shuy (2008) employed in analyzing the alert lexicon in product warnings based on Global Harmonization Standard (2013) and American National Standard Institute's (2002) legal yardsticks on warning the consumers.

The adjectives used in product warnings were also examined based on Marza's (2011) evaluative approach in analyzing the attributive descriptive words. On one hand, adverbs were analyzed according to Frey and Pittner (1999), Pitner (1999, 2000a, 2004) and Frey's (2000) usage of manner adverbs. Temporal adverbs were also explored based on Kiefer's (2007) framework on the 'time point' of adverbs. These linguistic tools aided the analytic procedural phase of examining the lexical attributes of Philippine consumer product warnings specifically the medicines.

2. METHOD

The study applied a textual evaluation of the linguistic features of medicine product warnings in the Philippines. By describing the lexical features of product manufacturers' word choice, frequency and percentage tools were applied.

2.1. The Corpus

Every medicine is believed to perform miracles for some groups of patients. Whether prescription or over the counter drugs, these medicines intend to save lives, enhance the patient's wellbeing, or provide them with hope that their health will be better upon medicine intake. Basically, mothers decide for their children on what medicine to take particularly in times of fever and flu. However, if the medicine is prescribed by doctors, mothers monitor the health condition of the family members.

The choice of medicines to be included as corpus was based on the survey conducted among 50 mothers in the supermarket and drug stores who purchased the medicine needs of their family. Below are the results.

Table 1 Surveyed Mothers' Top Medicine Needs

Products	Frequency	Percentage
Paracetamol	25	50
Ibuprofen-Paracetamol	10	20
Mefenamic Acid	5	10
Diatabs	3	6
Amoxicillin (liquid)	3	6
Alcohol	2	4
Aqua Oxinada	2	4
Total	50	100

Considering the ethical aspects of the study, the product brands and company names of medicine manufacturers were coded. Each product warning was masked; thus, Med was referred to medicines, while #1 (and so on) was assigned to each product based on the arrangement of warnings in the analysis of the research corpus. For specificity, clarity, and precision of the discussion, this paper selected and employed 50 warnings of medicine products which served as the major corpus of this construct.

2.1.2. Data Analysis

This research is anchored on the study of Shuy (1990, 2008) on the warning adequacy issues of product's cautionary texts. He presented several examples of linguistic consultations in civil cases that describes the theories and techniques applied by linguists in examining language evidence.

Qualitative method of research in examining the adequacy of medicine product warnings in the mainstream Philippine market, cautionary text was examined based on lexical features such as signal words, nouns, synthetic pronouns, field continuum, adjectives, and adverbs. The adequacy issues of product warnings were given importance, specifically in promoting comprehension alongside consumer safety. In processing the data, frequency and percentage counts were computed. Since the objective of the study was to ascertain the lexical features of product warning texts, the researcher did not apply some complicated statistical tools.

3. RESULTS AND DISCUSSION

Lexical features are words that contain distilled knowledge about the relationship between a particular communicative intent and its reception (Fussell & Kreuz, 1998). The study presents the lexical modes which encompass words or phrases in product warnings specifically on the existing safety texts of medicines.

3.1. Signal Words

In warning the consumers, an alert lexicon (Shuy, 2008) is placed before the main text of the product warnings to catch the attention of the product users. Global Harmonization Standards (2013) and ANSI (2002) recommend the terms DANGER, WARNING, and CAUTION to determine the degree of product hazards from highest to lowest. In identifying the degree of the hazard's gravity, ANSI (2002) designated three color-coded signal words to alert the consumers:

Defining and distinguishing the words "danger," "warning," and "caution," with "hazard" serving as a general cover term for the other three, as follows: DANGER indicates an imminently hazardous situation which, if not avoided, will result in death or serious injury. WARNING indicates a potentially hazardous situation which, if not avoided, could result in death or serious injury. CAUTION indicates a hazardous situation which, if not avoided, may result in minor or moderate injury. CAUTION may also be used to alert against unsafe practices. The three signifiers vary in the usage of modals will, could and may. The table below presents the signal words examined in the study.

Table 2 Signal words used in medicine product warnings

Signal Word	Frequency	Percentage
Precaution	22	44
Caution	17	34
Warning	7	14
Warning and Allergy	1	2
Stomach Bleeding	1	2
Important	1	2
Poison	1	2
Total	50	100

Based on the analysis, common among medicine product warnings in the Philippines utilized alert lexicons like warning, caution, and precaution such as the following:

Precautions: Keep away from eyes and mucous membrane. Keep away from Children. (MED#13)

Precautions: Always keep on container tightly close. Store at temperature not exceeding 30'C. (MED#16)

PRECAUTION attains 44 percent contribution in the corpus. Oxford dictionary (2016) defines 'precaution' as a situation taken beforehand to ward danger or secure safety. The sample extract

(MED#13) is giving emphasis on the location where it should be placed before and during its usage which is parallel to the objective of the word Precautionary. Though there is no specific directive in the use of the product, manufacturers intend to catch the attention of the consumers prior to their use of the medicines.

Another signal word is identified in the corpus, the utilization of CAUTION.

Caution: For external use only. Avoid contact with eyes. Prolonged used is discouraged. (MED#49)

Caution: If redness, irritation occurs, discontinue using and in case deep or punctured wounds or serious burns, consult a physician. (MED#17)

Signal word 'caution' transpired 34 percent in the corpus. However, extract MED#49 tells about what to avoid in using the product which is not related with the supposed content of the product warning extract (MED#17), it is symmetrical with the warning content of the previous. On the other hand, it varies in the last phrase recommending to consult a physician in case irritation occurs. Though the extract mentions possible risks, it does not state the nature of danger associated in the product, hence it does not correspond with the use of CAUTION as alert lexicon.

Meanwhile, alert lexicon such as 'warning' incurred 12 percent contribution in the corpus. It is used if the product will cause death to its users. However, extract MED#35 represents the mismatch between the warning content and its signal word.

Warning: Enzymes should not be used together with this solution. Do not boil internal in SEPTOCARE solution. (MED#35)

The usage of 'warning' according to ANSI (2012) and Global Harmonization Standard (UN, 2013) will lead the product user to death, however, it is in contrary with the content of the sample extract. If error in the usage of product persists, there is no warning of death specified in the main cautionary texts.

The most common alert lexicon used in the warnings was precaution obtaining 44 percent of occurrences. It was followed by caution with 34 percent and the warning lexicon with 12 percent. However, mix-up of signals words were also identified which incurred limited percentages in the corpus.

Pointing the variation of signal words applied on medicines, it was revealed that there were no standards in the usage of signal word in the existing product warnings. The identified mixed-up of signal words reflects the country's unspecified criteria on the use of signal word. According to Shuy (2008) signal word is the consumer's first line of defense, if the product liability law in the Philippines will not resolve the mismatch issue of cautioning the consumers, it will lead to confusion and will create a sense of business irregularities.

3.2. Order of Nouns

Generally, nouns and noun phrases are significant and purposeful parts of speech which determine product warning issues specifically on named identification. According to Viglioccoa (2011), nouns serve as the subject of discourse be it an object or a person; hence, the nouns of product warnings are categorized and underlined to mark the observations.

3.2.1. Concrete and Abstract Nouns

Concrete and abstract nouns are directly linked to perceivability. Concrete nouns are known for physical entities with characteristics like shapes, parts, materials, and alike whereas abstract nouns lack physical attributes (Katja, 2008; Crystal, 1995).

In analyzing the corpus, Lakoff's (1987) Cognitive Linguistics underscores the ontology of noun entities was applied, specifically Lyons's (1977) peculiarity of the first three orders of nouns. The summary of results is shown in Table 3.

Table 3 Order of Noun used in medicine product warnings

Order of Nouns	Frequency	Percentage
First	9	18
Second	20	40
Third	7	14
Total	36	72

Defined by Lyon (1977), first order entities are directly referring to a person, animals, objects, and other organisms which are situated in space. The current research categorized the concrete nouns of medicine product warnings.

Precautions: Always keep on container tightly close. Store at temperature not exceeding

30'C. (MED#16)

The study found out that 18 percent of product warnings utilized first-order entity. The words *eyes* and *nostrils* were employed to directly inform the patients about the body parts which might be prone to danger. Meanwhile, analgesic was specified to medicine which might cause adverse reaction upon intake, and container named the appropriate place where the medicines should be kept. Such usage implied that medicine manufacturers were trying to be specific in their warnings and they are attempting to create a 'temporary concept-formation' (Lyons, 1977). However, 18 percent of occurrences were very minimal.

Meanwhile, abstract nouns bring up the second-order and third-order units. Lyon (1977) highlights that second order entities involve events, processes, and state of affairs. Below is the extract.

Caution: Should be used with caution in patients with hypertension, in patients whose cardiac reserve is poor, and those with heart failure since deterioration of heart failure has been noted. (MED#12)

The terms hypertension and deterioration connote state of affairs as to illness and worsening condition of heart might take effect if medicine will not be taken with caution. Second order entities are definitely observable and perceptible. Surfacing these nouns in medicine warnings occurred 40 percent in the corpus, hence, naming activities which might take place once the warnings will be neglected by the consumers of pharmaceutical drugs is evident.

On one hand, third order entities are abstract items such as concepts, propositions, or more generally ideas outside place and time. Consider the extract below.

Caution: Before taking bisacodyl, tell your doctor or pharmacist if you are allergic to it; or if you have any other allergies. This product may contain inactive ingredients, which can cause allergic reactions or other problems. Talk to your pharmacist for more details. (MED#3)

Third order entities obtained 14 percent of the occurrences in the corpus. Nouns such as any other allergies and inactive ingredients are mental phenomena which the former means another type of allergy while the latter connote unknown substance which Mackenzie (2008) emphasizes that these nouns are unobservable. Lyon (1977) highlights that third-order entities can be asserted or denied, remembered or forgotten. Such activities are considered an austere 'no-no' in product warnings.

3.2.2. Pronoun

Traditional grammar posits a category of pronoun to denote a class of words which are said to 'stand in place' (the meaning of the prefix pro) or 'refer back to' noun expressions (Radford, 2004). The second person pronoun YOU and possessive pronoun YOUR are used generically, referring to the warning readers or product consumers in general. Indefinitely, YOU and YOUR pronouns signal a direct and personal address to the consumers. According to Kaur et.al (2013), this dealing of people on an individual or particularized basis is referred to as 'synthetic personalization' (Fairclough, 1989; Kaur et.al, 2013), a method of addressing mass audience (consumers) as though they were individuals through inclusive language usage.

The study revealed that there was an occasional use of direct address in medicine products as this is manifested in the use of second-person personal pronoun. Consider the extracts.

> Before taking bisacodyl, tell your doctor or pharmacist if you are allergic to it; or if you have any other allergies. This product may contain inactive ingredients, which can cause allergic reactions or other problems. Talk to your pharmacist for more details. (MED#3)

If you consume 3 or more alcoholic drinks every day, ask your doctor whether you should take acetaminophen or other pain relievers/fever reducers. (MED#5)

Five or 10 percent among 50 medicines employed direct addressing to consumers. Communicatively, these five product warnings desire to establish an interpersonal relationship with the readers, one way of articulating the message of the warnings. In the end, pronoun YOU serves as generic address to the product consumers or users, thereby an informal type of speech and writing.

3.2.3. Words in Field Continuum

Field highlights what the text is about (Donnell, 2010) and such typical fields could be classified as science, education, war, medicine, sports and others. To be able to describe the texts of product warnings, Halliday's field continuum (Halliday and Matthiessen, 2004) is employed as a guide. The specialized scale is classified as: everyday word, specialized language, and highly technical.

Table 4 Words in field continuum used in medicine product warnings

Words in Field Continuum	Frequency	Percentage
Everyday	15	30
Specialized	13	26
Highly Technical	22	44
Total	50	100

3.2.3.1. Everyday Language

Everyday language uses ordinary and familiar words which a typical consumer can easily understand. Words are widely known and frequently used as they tend to refer to concrete things rather than abstract ideas. The corpus incurred 30 percent of occurrences in the warnings.

Precautions: For external use only. Avoid contact with eyes. Do not swallow. Keep away from heat, sparks and flame. Keep container tightly closed. Keep out of reach of children (MED#46)

The results entailed that manufacturers of medicines tried to communicate with the product users in a comprehensible language. However, the limited number of percentage did not suffice the language simplicity of existing warnings.

3.2.3.2. Specialized Vocabulary

Specialized word or vocabulary is considered from a broader perspective, specifying that it is a technical terminology and semi-technical vocabulary (Rizzo, 2013; Hyland, 2007; Nation, 2001; Alcaraz, 2000). It is made up of 'lexical units of various levels of specialization' (Rizzo, 2013), or a general language which acquires a specialized meaning in the discipline.

If accidentally swallowed induce vomiting and call a physician. (MED#1) Amlodipine besilate should be used with caution in patients with hypertension, in patients whose cardiac reserve is poor, and those with heart failure since deterioration of heart failure has been noted. (MED#12)

Medically, the word 'induce' means to cause (something) to happen or exist as to give a (pregnant woman) special medicine in order for her to give birth. Hypertension, on one hand, refers to the abnormality of blood pressure resulting to arterial blood pressure otherwise known as high blood. Such vocabulary incurred 26 percent of occurrences in the existing warnings.

The results were in the contrary of Shuy's (2002) call for the comprehensibility of product warnings. The consumers' limited knowledge on words and word retention for specialized vocabulary will become their dilemma not only in understanding the safety information but also in taking appropriate action in times of emergency.

3.2.3.3. Highly Technical

Product warnings are made available to the public; however, the identification of jargons among the cautionary texts created a mismatch in the trading context since they are considered to be a hybrid language. Highly technical words incurred 44 percent in the existing corpus.

Precaution: Citrimoxazole should not be given to patients with a history of sensitivity to it or to the sulfonamides or trimethoprim, and to infants below 6 weeks of age because of the risk of Kirnecterus from sulfonamide component. (MED#9) Precaution: It should be given with caution to patients with glaucoma, cardiovascular disorder, diabetes mellitus, hyperthyroidism, hypertension, urinary retention, prostate hyperplasia, or pyloproduodenal obstruction. (MED#23)

The common public's level of awareness about science and technology seems to be alarmingly low (Eurobarometer 1993; Miller 2000), hence, the inclusion of jargons in product warnings manifests a great disparity about the language and comprehension abilities of product consumers. Furthermore, jargons create the effect of making the warning readers feel eliminated and alienated (Halliday, 1989) from using products.

Based on the results, the existing dangers are difficult to comprehend as reflected by 48 percent of vocabulary in the existing warnings containing the use of highly technical language. This goes against the Global Harmonization Standards (UN, 2013) and Consumer Act of the Philippines (1992) citation on comprehensibility in notifying the product users of existing dangers brought by medicines.

According to the study of Nguyen (2011), the use of medical jargon leads to poor communication between product manufacturer and product consumer, and consequently, leads to ineffective cautionary care. Philippine product warnings manifest that these terms are fertile ground for everything from a funny consumer misinterpretation to a life-threatening medical error.

3.2.4. Evaluative Adjectives

Highlighted by Leech (1989), adjectives are the leading open word class in English after nouns and verbs. Grammatically and semantically, they carried the same level of vitality as the other content words in the linguistic code. Adjectives of medicine product warnings are evaluated since these are responsible for classifying events or entities (Marzá, 2011) or simply by depicting their qualities.

Evaluative adjectives are among the noteworthy means of examination as the manufacturers point out what product and warning qualities should be avoided by the consumers. For precision purposes, this paper examined the adjectives' evaluative positions as to attributive.

Attributive

Attributive adjectives are positioned before the nouns (Marza, 2011). The extract signifies its worth in the warning discourse.

Cautions: If redness, irritation occurs, discontinue using and in case deep or punctured wounds or serious burns, consult a physician. (MED#17)

Identifying 18 percent attributive adjectives in the corpus, it was disclosed that this type of adjective will significantly persuade the consumers to keep away from product risks. It aims to

help the product users in carrying out safe behavior. However, its minimal inclusion in the warning texts provides a lesser evidence of risks among consumers.

3.2.5. Adverbs

According to Geuder (2002), adverbs refer to adverbial modifiers which are morphologically derived from an adjectival base, or are formally identical to adjectives. It is a word that gives an explanation of where, when and how an activity or event occurred.

3.2.5.1. Manner

Constructed by adding ly to adjectives, manner adverbs drawn its striking property as degree words (Abeille, 2003). These may be characterised by the informal communication, however, they immediately instruct the consumers on 'what to do' before using the medicines. Consider the extract below.

Caution: Do not swallow. Consult a physician immediately if accidentally swallowed. Do not apply in or near eyes. Not applicable for deep wounds. Keep tightly closed and store at temperatures not exceeding 30. Keep out of reach of children. (MED#11)

From the analysis of data, manner adverbs incurred 12 percent appearance in the corpus which highlights that the conveyed information has the 'integrated' prosody (Abeille, 2003) and have given the product users the emotional hint and the 'intonationally' (p. 29) about the manufacturers desire for the consumers to take necessary action (immediately, tightly) in case of emergency. This promotes the sense of urgency among product warnings which consumers should follow.

3.2.5.2. Temporal

Adverb of time is linked whenever an event may occur and tells how long an incident lasted as presented in the following extracts:

Caution: Should be given with care to patients with impaired kidney or liver functions. Chronic use should be avoided. Daily use may potentiate oral anti-coagulant. (MED#2) Caution: Alcohol warning: if you consume 3 or more alcoholic drinks every day, ask your doctor whether you should take acetaminophen or other pain relievers/fever reducers. Acetaminophen may cause liver damage. (MED#5)

Since time reference is enlisted in the warnings, temporal adverbials serve to mark discourse segment boundaries of medicine safety information. Based on MED#5 extract, it serves as instructions to the readers on how long an event or state endures (Nguyen, 2010).
In addition, temporal adverbs like *after, before,* and *always* are used to indicate when something happens in the past, present or future.

Warning: Shake the reconstituted suspension before using. Food drugs devices and cosmetics act prohibits dispensing without prescription. Keep in dry place, store below 25. (MED#36)

Based on the extract, before means 'before that moment' (use) occurs. Hence, there is a need to read and follow the stated directives. In this regard, the outcome considers the common usage of indefinite adverbials, where they do not count the particular date that the event occurred on, but only a particular property of when the event takes place (Spejewski, 1996). If an adverbial does

not fit the constraints on the reference time, then it cannot be used to modify the 'reference time' (p.263).

Significantly, the inclusion of temporality in product warnings reminds the consumers about the exact time frame when a medicine should be taken or avoided, hence following instructions on, before, or during the utilization of the product are specified.

Importantly, temporal adverb specifies point or boundary in time to whom an event occurs or lasts. Consider the extract below.

Citrimoxazole should not be given to patients with a history of sensitivity to it or to the sulfonamides or trimethoprim, and to infants below 6 weeks of age. (MED#9)

This time point refers to the duration on how an eventuality will take place to whom Spejewski, (1996). It seems to introduce a new reference time to the consumers.

With the minimal instances of adverbs of manner and time which respectively incurred 14 and 18 percent in the corpus, it empirically shows that adverbs are slightly incorporated by manufacturers in writing the warnings of medicines. The minimal practice in explaining where, when and how an accident or erroneous event will occur posits limited reference in the course of product usage.

3.2.6. Modals

Modality is a category of linguistic meaning which necessitates the expression of possibility and necessity. Modals are very evident among product warnings and to determine the degree of the possibilities. This paper used Halliday's tenor continuum (1988) as a guide in analyzing the extracts. The table below showcases the occurrence of modals in the existing warnings of medicines.

Table 5 Modals used in Medicine Product Warnings

Modals	Frequency	Percentage
High	11	22
Medium	20	40
Low	19	38
Total	50	100

The modal MAY occurred 19 times or 38 percent in the warning discourse. The Writing Center of North Carolina (2013) explained that the use of modal MAY weakens the certainty of a sentence. Since *may* belongs to the lowest level of continuum, it manifests a weaker possibility that hazards might occur. The extract below is adapted for analysis.

Precautions: One ingredient in this product is acetaminophen. Taking too much acetaminophen may cause serious (possibly fatal) liver disease. Adults should not take more than 4000 ml. (4 grams) of acetaminophen per day. Daily use of alcohol, especially when combined with acetaminophen, may increase your risk for liver damage. Avoid alcohol. (MED#22)

Caution: Alcohol warning: if you consume 3 or more alcoholic drinks every day, ask your doctor whether you should take acetaminophen or other pain relievers/fever reducers. Acetaminophen may cause liver damage. (MED#5)
Though may can also express irrelevance in spite of certain or likely truth, it also produces a contradictory effect associated in using the product e.g. may cause liver damage articulates a serious health effect which requires a stronger modal.

The result is similar to the findings of The New Mexico Court (Malik and Tiersma, 2013) which specified a medication case in New Mexico, where the phrase 'it may damage the kidneys' was written in very small letters. The court held this statement as too vague and misleading which should have straightly informed the purchasers that 'it will damage the kidneys'. This creates a direct communication practice between the manufactures and the consumers.

Since the function of the modals is to help the consumers in determining the possible effects and the necessary steps to be done in case of emergency, the use of 'may' may lead the patients in uncertainty specifically in determining the product risks.

Meanwhile, another essential finding is the emergence of medium level modals. Below is the extract.
Precautions: Contraindicated in patients known to be sensitive in penicillin. It should be used with caution in patients with known history of allergy to penicillin V.

The usage of should incurred 40 percent uses in the corpus which entailed that product manufacturers have their strong desire in warning the consumers. Thus, strong compliance should be followed by the product users since there is a higher degree of risks that may emerge in the process of product usage. However, based on the corpus it failed to mention the nature of the hazard and possible consequences that might occur once the product is misused.

4. CONCLUSION AND RECOMMENDATION

The conduct of analyzing the lexical features of medicine product warnings manifest the communicative intent of product manufacturers specifically on helping the consumers in doing and following the product precautionary measures. It can be supposed that linguistic features such as the lexical aspects bring a vital impact on the legal-content adequacy of medicine product warnings, thus, helping the consumers in coming out with informed decisions during the presale and post sale of the medicine products.

Considering the informational tidbits from the different medicine warning extracts, Philippine legislatures can further improve the existing product warning law through Consumer Act of the Philippines based on various research and issues associated with consumer product warnings. It brings to fruition the application if not introduction of parallel guidelines in Philippine context. Significantly, the law can set the standard of reasonableness for the business industry to follow, thus upholding the consumers' right to live in a safe and healthy milieu.

REFERENCE

[1] Abeille, A., & Godard D. (2003). The syntactic flexibility of adverbs: French degree adverbs. Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar, Michigan State University (pp. 26-46). Stanford, CA: CSLI Publications.
[2] American National Standard Institute. (2002). American national standard criteriafor safety symbols. VA: National Electrical Manufacturers Association.
[3] Fairclough, N. (1989). Language and power. United Kingdom: Longman.

[4] Food and Drug Administration Philippines. (2012). Declaring the identified medicine and cosmetic products to be imminently injurious, unsafe or dangerous.Retrieved from http://www.fda.gov.ph/attachments/article/17138/FC2012-009.pdf
[5] Frey, W., & Pittner, K. (1999). Adverbial positions in Deutsch-English compared. Studia Grammatica, (48), 14-40.
[6] Globally Harmonized System of Classification and Labelling of Chemicals (2013). United Nations. Retrieved from https://www.unece.org/trans/danger/publi/ghs/ghs_rev05/05files_e.html
[7] Halliday, M.A.K. (1988). On the ineffability of grammatical Categories. In J.J.Webster (Ed.), On grammar: Collected works of M.A.K. Halliday (pp. 291-322).London and New York: Continuum.
[8] Halliday, M. and Matthiessen, C., (2004). An introduction to functional grammar. 2nd ed. London: Arnold. p.121.
[9] Heaps, C.M., & Henley, T.B. (1999). Language matters: Wording considerations in hazard perception and warning comprehension. Journal of Psychology: Interdisciplinary and Applied, 133, 341-351.
[10] Hancock, H., Fisk, A., & Rogers, W. (2002). Comprehending product warning information: Age-related effects and the roles of memory, inferencing, and knowledge.Human Factors.
[11] Kaur, K., Arumugam, N., & Yunus, N.M. (2013). Beauty product advertisements: A critical discourse analysis. Asian Social Science, 9(3), 61-70. retrieved from doi:10.5539/ass.v9n3p61
[12] Lakoff, G. (1987). Women, fire and dangerous things. Chicago: University of Chicago Press.
[13] Lyons, J. (1977). Semantics part 2. London and New York: Cambridge University Press.
[14] Mackenzie, J.L, (2008). The contrast between pronoun position in European Portuguese and Castilian Spanish: an application of Functional Grammar. Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives. Amsterdam & Philadelphia PA: Benjamins. p. 51-75.
[15] Malik, N. (2010). Product warnings: Language, law, and comprehensibility. Journal of Literature, Culture and Media Studies, 2(10), 102-106.
[16] Marzá, N.E. (2011). A comprehensive corpus-based study of the use of evaluative adjectives in promotional hotel websites. Odisea, 12, 97-123.
[17] Miller, J. (2008). Otherness. The SAGE encyclopedia of qualitative research methods. Thousands Oaks, CA: Sage Publication.
[18] Shuy, R.W. (1990). Warning labels: Language, law, and comprehensibility. American Speech, 65, 291-303.
[19] Shuy, R.W. (2005). Creating language crimes. New York: Oxford University Press.
[20] Shuy, R.W. (2008). Fighting over words: Language and civil law cases. NewYork, USA: Oxford University Press.
[21] Spejewski, B. (1996). Temporal subordination and the English perfect. In T. Galloway, & J. Spence (Eds.), SALT VI (pp. 261-278). Ithaca, New York: Cornell University.
[22] Tiersma, P., & Solan, L. (2002). The Linguist on the Witness Stand: Forensic Linguistics in American Courts. Linguistic Society of America, 78 (2), 221-239.
[23] Wogalter, M. S., (2006). Handbook of Warnings. Ergonomics in Design: The Quarterly of Human Factors Applications. Mahwah, NJ: Lawrence Erlbaum Associates, 78 (2).

AUTHOR

Shielanie Soriano-Dacumos is Associate Professor at the University of Rizal System Binangonan Campus where she led the then General Education Center and the Student Publication and Mass Media. She completed her Doctor of Philosophy in English Language Studies at the University of Santo Tomas, Philippines with her research coring on the Linguistic Features, Adequacy, Comprehensibility, and Readability of Philippine Consumer Product Warnings, a pioneer study in Forensic Linguistics in the Philippines.

A DESKTOP APPLICATION TO HELP SPEAKERS SWITCH SLIDES BY USING AI AND VOICE RECOGNITION

Yixin Liang[1], Marisabel Chang[2]

[1]Portola High School, Portola High School, 1001 Cadence, Irvine, CA 92618

[2]Computer Science Department, California State Polytechnic University, Pomona, CA 91768

ABSTRACT

Presentation is a skill that everyone has, and it is very commonly seen in companies, schools, conferences, etc [1]. And the purpose of a slide is to give the audience a better understanding of the topic and to add ideas that they forgot to mention [2]. It also adds visual support to the speaker's discussion. Usually the presenter held a slide remote or just used their computer to control the slide pace while presenting. However, the slide remote can often be unstable due to battery switching. Even those who do not have a slide remote are unable to ensure a smooth presentation because they need to constantly switch back and forth on the computer screen with the mouse, which not only makes the speaker more nervous but also likely to skip the slide. Slidecontroller uses existing AI technology,voice recognition, as a medium to allow users to enter the transition word used to switch slides [3]. For example, when the user enters "Now I am going to talk about" when this word is spoken the Slidecontroller will receive the voice and match the speaker's turn to the next slide. The user can be creative with the keyword selection that best fits their presentation vibe. Or the user could use the Slidecontroller default option which controls the slide by simply saying "Next" to go to the next slide, "Previous" to go to the previous slide, and "Thank you" to stop the App to prevent from catching a similar keyword that accidentally switches the slide [4].

KEYWORDS

AI, voice recognition, Slide controller

1. INTRODUCTION

In modern society, almost all successful entrepreneurs are presenters [5]. Whether it is entrepreneurs or the workplace daily work report, results can not be separated from speech, over time has become a necessary skill for everyone. And the speech itself is a skill to help put this realization of self-worth and create value for others to open up the road, able to enhance the connection between each other and maximize the output of their own experience. And to further assist in improving people's presentation skills, the first thing to address is technology [6]. Especially on some formal occasions with the mouse, slide remote, or keyboard constantly switched or very distracting behavior. To allow users to reduce the anxiety of the presentation, and in the case of not considering the use of slide remote and keyboard, the speaker's hands can be free to do gestures to increase the interest of the speech. At the same time, there is no need to worry about the system short circuit because only a computer and the presenter's "voice" is needed. This not only improves the efficiency of the presentation but also makes the transition between slides more smooth [7]. Slidecontroller also takes into account the inconvenience of the

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 151-160, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130412

disability group, so it uses only the user's voice to make this desktop application more convenient and comprehensive.

Currently, two technologies can control slide conversion, one is the traditional hardware configuration of slide remote, and the other is the new google slide presentation remote on cell phones launched by google in 2018. First of all, let's analyze the traditional slide remote, let'stake the Canon PR10-G wireless presentation remote as an example to analyze the object. Same as with other slide remotes, the speaker needs to hold it in his/her hands during the presentation. The difference is that the Canon PR10-G wireless presentation remote is designed with injection-molded plastic for easy grip, but the location of the buttons will inevitably be worn and cause the remote to malfunction under prolonged use. And some speakers are not comfortable with holding something in their hands while presenting, and this only makes them more nervous. Second thing to analyze is the core of each presentation remote control, the control that allows the user to control the slide presentation at any time. On top of the Canon PR10-G, the wireless presentation remote is the typical forward and backward buttons, and below it is the "present" button. This can be used with PowerPoint or Keynote to enlarge images, videos, or charts to full size. The problem is that the sensitivity of the controls is unquestionable, so users are likely to press the forward or back button more than once in the most stressful situations, causing the slide show to miss. The last is its battery life, Canon PR10-G wireless presentation remote control requires users to replace the battery, and can not be recharged, and every time to replace the battery is also very environmentally unfriendly behavior, especially teachers in the use of 7 days a week its battery life is only 7-9 months, so the average consumption of the battery is at least two per year. The next existing technology to analyze is google's new cell phone presentation remote control, the basic function is almost the same as the traditional slide remote, just replaced by holding the phone in your hand only. Google developed the cell phone presentation remote control's biggest drawback is that it only applies to slides made with Google Slide and others like PowerPoint. can't be used. And it becomes very inconvenient and unattractive on serious occasions or when the use of cell phones is not allowed.

Slidecontroller is a desktop application based on AI voice recognition technology [8]. By accurately receiving the user's voice, it collects data and analyzes the hidden keywords, and connects them to the "up arrow" and "down arrow" on the computer keyboard to make relative commands. The goal of this application is to allow all users to speak without worrying about any technical problems. Even if the speaker can not pay the price of a traditional slide remote, then this app must be the most affordable and accurate replacement. In contrast to existing remote control technology, Slidecontroller allows users to customize their Keywords and works with all Slide-creating apps, not just Google slide or PowerPoint. The Slidecontroller is also equipped with a default option if the user just wants a small presentation. The technology used for this option is Hotword detection, similar to the technology used by Siri or Alexa, which is a miniature algorithm that monitors the audio stream of special hotwords [9]. Slidecontroller is very easy to use and does not require the user to have anything in their hands while presenting, so anyone can use it.

The best proof of the usefulness of Slidecontroller desktop applications on the spot is its use in everyday presentations. As the developer of the desktop application myself, I use Slidecontroller for every presentation opportunity at school, and the actual results are very good, helping me to relieve my nervousness to a greater extent and adding more hand movements to make the presentation more vivid. The data proves that Slidecontroller has completely replaced the existing slide remote technology with efficiency and accuracy. The number of seconds it takes to rotate the slides after each keyword is detected by the record. The accuracy of Voice recognition was also checked to ensure that the correct keywords were collected. After nearly two thousand

tests, the Response time was only increased to about 2.3 seconds when the user Customized more than 5 keywords, but all other functions were stable and accurate.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that were found while developing the Software and finding the most accurate Voice Recognition AI library; Section 3 focuses on the component that were used to solve the challenges as mentioned before and will present part of the code to shows the details of how the AI technology was used; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5; Section 6 gives the concluding remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Efficiency & Accuracy

The core of Slidecontroller, a desktop application, is the clever use of existing AI voice recognition technology to connect to the front and back keys of the keyboard for control. Therefore, the voice recognition library needs to be carefully selected, and the most basic requirement is that it should be able to run on both Windows and IOS platforms without any problems. Then is the accuracy of the received speech. In the Slidecontroller system accuracy is very important, because the computer needs to store the correct data to output the right command, but the existing AI speech recognition technology is limited in inclusiveness [10]. For example, when the user's voice input with another country's accent, voice recognition will recognize the wrong word. The accuracy of speech recognition is also affected when the user maintains a slight pause of more than six or seven seconds. Only when the user speaks perfect English and there is no pause, the speech recognition can be 100% error-free. But because the current AI speech recognition technology has not been able to expand its inclusiveness, the space available is limited. The next problem is the efficiency affected by the accuracy of speech recognition. Without the correct recognition of the user's voice output, Slidecontroller is unable to complete the command of switching slides the first time. The existence of Slidecontroller is meaningless if it cannot respond to the user's needs in the first place, and this is the biggest challenge of this desktop application.

2.2. Voice Receiving

As mentioned earlier, Slidecontroller has very high requirements for the speech recognition library, so the reception of the user's voice needs to be very subtle and not affected by outside influences. In most cases, there will be an audience on the stage during the speech, and if there isa little noise from the audience, speech recognition is likely to be affected, causing the stored content to be different from what the speaker describes. Or when the speaker is at a certain distance from the computer, speech recognition is difficult to capture complete sentences, and all these factors can directly affect the operation of the Slidecontroller. But Hotword Detection, which is used for the default keyword option of Slidecontroller, achieves the maximum absolute sensitivity of speech recognition [11]. But unlike normal speech recognition libraries, Hotword detection is limited to one or two keywords first, to improve the detection speed. But to maintain the diversity of Slidecontroller, for users who want to create their keywords, then the regular speech recognition library is inevitable. Therefore, the speech recognition library needs to be sensitive and unaffected by any external factors to maximize the benefits of Slidecontroller

3. SOLUTION

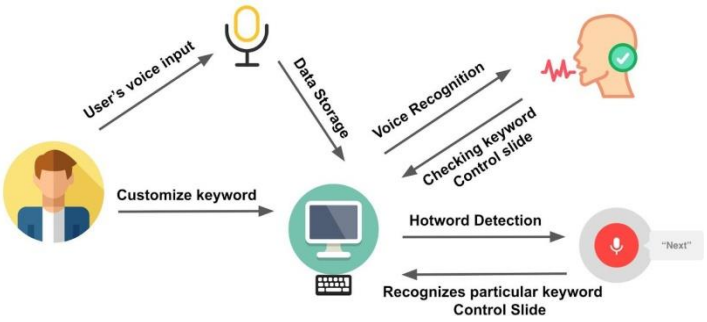


Figure 1. Overview of the solution

As the above prototype shows, the main source of data for Slidecontroller is the user's audio source. The computer's microphone picks up the audio and then the computer stores the data, which is analyzed by different algorithms depending on the keyword option selected by the user. The first important step is the computer keyboard control, which converts the control keys that can control the slide transition into commands. Then is the Hotword detection, where the user will use the default keyword, and the last one is if the user chooses to customize his keyword, where the Slidecontroller will automatically switch to Voice Recognition mode. Combining the above three main sections with the post-processed GUI screen, the user can operate more conveniently and smoothly. Figure 2 shows the complete Slidecontroller in the user's view. If the user chooses the default option, then first chooses whether the system is Windows or IOS, Picovice will change the Hotword Detection template according to the computer platforms but the function is still the same [12]. Then the user needs to copy the link at the top of the slide to the text entry of "Insert Link". Then if the user wants to save the link, they can click "Save" and name the link to make it easier to find. Finally, the user can control the slideshow by saying "Next" and "Previous". There is also a detailed tutorial in the lower right corner, which can be accessed by the single question mark icon. If the user wants to change the mode, click the arrow below to switch to Customize keyword, the procedure is similar to Default Keyword mode, you don't need to worry about your computer system, just copy the slideshow link to the corresponding Text entry and enter the corresponding Customize keyword to control the slideshow. After all the selections, the user can then click "Start Presentation" to begin!

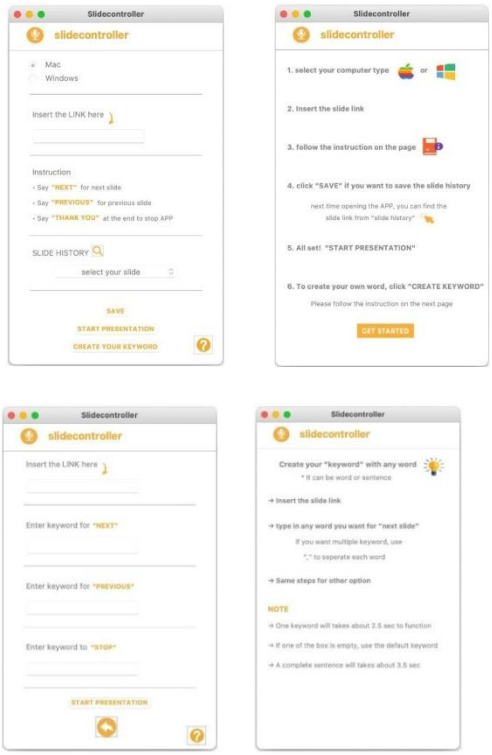


Figure 2. Overall looking of the desktop application

```
pageNumber = 1 # keep track on which slide
def nextPage():
    global pageNumber
    keyboard.press(Key.page_down)
    keyboard.release(Key.page_down)
    pageNumber += 1
    print("next function is working properly")
```

Figure 3. Page number code

As mentioned earlier Slidecontroller is composed of three main components. And the very first step is to control the slides using the up and down keys of the computer keyboard. Generally, in the absence of a slide remote, the presenters can only use the up and down keys of the computer for control, and the slide remote itself uses this method. So, as shown in the following code, the "Next" function for switching to the next slide is used as an example, and the first step is to define the “pageNumber” accumulator to keep track of the slide number. An accumulator is being defined, and then using the keyboard press function in the keyboard module is to let the keyboard autonomously click the forward button. The keyboard simulator is the basis of the whole Slidecontroller, and the final product is made up by adding voice control to the keyboard control.

```
def open_slide():
    global pv
    if urlEntry.get() == "":
        url = slide_history[value_inside.get()]
        webbrowser.open(url)
        time.sleep(3.5)
        present()
    else:
        webbrowser.open(urlEntry.get())
        time.sleep(3.5)
        present()
    if pv == True:
        voice3()
```

Figure 4. Default option code

The second component is the Default option, which is created using Hotword detection technology and covers the "Next" keyword to turn to the next slide, the “Previous”keyword togo back a slide, and the “ Thank you ” keyword at the end of the presentation to stop the program. Hotword detection is very similar to the Always-Listening Commands technique,which executes commands by using multiple words with the help of Always-Listening Commands. The difference is that Hotword Detection relies on hotwords, trigger words, keywords, or wake-up words to activate the dormant software. It is a combination of voice activation and Always Listening Commands. Also, voice activation is a key point of Hotword Detection, which activates the application by voice and then works with Always-Listening Commands to control the slideshow. Then by combining the previously set keyboard controls with Hotword Detection, the slideshow can be controlled by touching the "Next", "Previous",and "Thank you" keywords to complete the slideshow control commands in time. As shown in the code above, a web browser opening technique was added to allow users to copy the slide URL and then the computer would automatically convert it to present mode, making it more convenient without the need for users to do it by hand and allowing the audience to better understand the content of the slide after zooming in.

```
def verify_keywords(self,text):
    text = text.lower().replace('.', '').replace(',', '')
    print(self.next)
    if text == self.next:
        print('next')
        nextPage()
    elif text == self.previous:
        print('previous')
        previousPage()

    elif text == self.thankyou:# stop the program
        print(self.thankyou)
        sys.exit(0)
```

Figure 5. Customize keyword code

The last component is the creative Customize Keyword, where users can use their imagination to create a keyword that fits the speech. Speech recognition technology uses Always-Listening technology, which always listens to the voice used and activates Speech-to-Text to transcribe the

surrounding dialogue when speech is detected. In more detail, the system first analyzes the audio, then breaks it down into parts, digitizes it into a computer-readable format, and finally uses an algorithm to match the audio with the most appropriate text representation, and then looks for the specified keyword in the matched text [13]. As the above coding shows, it first confirms what the user specified as the keyword and then works with the keyboard to the command is completed with the keyboard control.

4. EXPERIMENT

4.1. Experiment 1

To test the efficiency and accuracy of speech recognition libraries supported by existing AI technology, I compare two speech recognition libraries, Google Voice Recognition and Assembly AI, using speech recognition filtered from 15 different companies before. I will apply each of the two voice recognition libraries to the Slidecontroller system, where the keywords will be the same for each voice input, and then record how many seconds it takes to respond and implement the instructions to switch slides.

Two separate computers with the same platforms(either both windows or ios). The speaker should stand at equal distance to both computers to ensure the computer receives the voice data equally. Run both Slidecontroller with different voice recognition libraries at the same time, and speak the same amount of words (constant variable). Different amounts of keywords will be tested 3 times, the maximum keywords included will be 5, so the data will take the average of 15 times. Then record the number of seconds it takes to process the keywords and respond to turnthe slide.

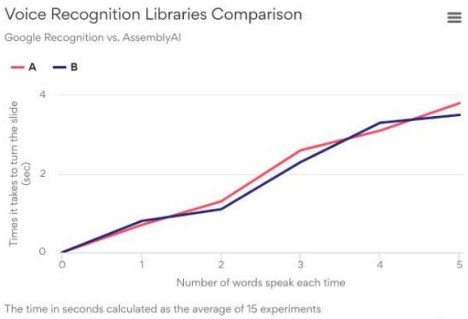


Figure 6. Voice recognition libraries comparison

As shown from the statistical chart above, the difference between Google Recognition and AssemblyAI is only a fraction of a second. But this fraction of a few seconds difference will directly affect the efficiency of the Slidecontroller. In the 15 tests, AssemblyAI is more stable and accurate than Google Recognition in terms of general trends. Google Recognition is a popular speech recognition system that does not require payment and is not used for software development, considering the target audience. Assembly AI, on the other hand, is a paid speech recognition library, but for a fee of ¥5, you can use it for an unlimited time. However, there is no difference between the two speech recognition libraries in general, but Slidecontroller requires higher sensitivity, so AssemblyAI is better and the cost is reasonable, so users do not have to worry about the financial burden.

Experiment 2

In this experiment, two algorithms that were used in the Slidecontroller are being compared based on time efficiency, which will show how many seconds it takes for the keyword to react. Hotword detection is provided by the picovoice platform, and speech recognition is provided by Assembly AI. The purpose of this test is to allow users to better visualize the operation and response time of the two different programs so that they can be more secure when choosing one. It is also to ensure that the user selects the appropriate function for the presentation. Since Picovoice provides hotword detection only with the Default keyword option provided by Slidecontroller (all the default keyword that use to control slide is one-word “Previous” and “Next”, except the ending keyword that is used to stop the App which is two words “Thank you”), it does not provide the ability for users to create their keywords. Assembly AI, however, can fulfill this need by using the freedom of Voice Recognition combined with the Keyword Capture feature to allow users to create keywords that match the atmosphere and fluency of the presentation.

Two separate computers with the same platforms (either both windows or ios). The speaker should stand at equal distance to both computers to ensure the computer receives the voice data equally. The Hotword detection will run the same keyword for 15 times, and the Assembly will run different keywords based on user customize keywords. Then record the number of seconds it takes to process the keywords and respond to turn the slide.

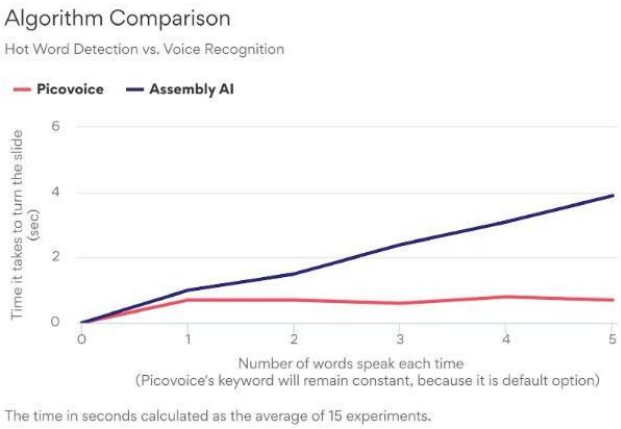


Figure 7. Algorithm comparison

As shown in the line data graph, the hotword detection provided by Picovoice has a response time of basically no more than one second when the user speaks the keyword of default and is very stable. Because of its special voice monitoring system, "Always-listening Commands" allows it to always be listening, knowing that the user speaks a specific keyword to activate the corresponding command. On the other hand, Assembly AI wins the comparison with Google Recognition, but the algorithm of detection is very different from Hotwork's, and the response time increases according to the number of keywords. Therefore, the results of this experiment are

intended to show users the advantages and disadvantages of the two different algorithms and to give them a better reference to evaluate which one to use. If the user is not overly concerned about the overall smoothness or does not have any idea to customize the keyword, it is recommended to use the default option keyword already provided by Slidecontroller, which not only guarantees overall accuracy and fast response. But if the user wants to incorporate more personal ideas or wants to fit the speech scene, you can use Assembly AI, although the responsiveness will be slightly inferior to Hotword detection, but comparable to other Voice Recognition with the same algorithm, and its accuracy can also be guaranteed.

In the existing AI speech recognition technology, Hotword Detection, and Assembly AI are used in much the same way. For Assembly AI, the more words the user input, the longer it will take. But according to the graph above, it shows the maximum amount of time for 5+ keywords is 3 seconds which doesn ’ t affect the overall presentation. However, Hotword Detection is recommended for less formal situations because the keywords that can be used are very limited. For example, students using Slidecontroller's default keyword when the teacher does not provide a Slide Remote is a good choice and will respond promptly. Both speech recognition libraries have very good performance and accuracy, so no matter which one you use, it will help you to complete a perfect speech.

5. RELATED WORK

In 2018, Google released "Remote for slides," a slide remote that could be operated by phone [14]. Each operation required a 6-digit unique number that was linked to the relevant slide that the user was presenting. Users can utilize this free web application by following a few straightforward steps. On the smartphone UI, there are simply two enormous buttons that read "Next Slide" and "Previous Slide," respectively. The Back and Forward buttons that come with Google Slide are combined in this app. In the first week after its introduction, it attracted a lot of attention, but its flaws soon became apparent. Users must have both a cell phone and a computer, without either one then the slide will not be able to turn into the present mode, this step must be operated on the computer. Also, if the user's phone is blacked out during the presentation, then it will directly lead to the APP not working. And on some formal occasions, holding a cell phone in your hand is not suitable and affects the audience's perception. Slidecontroller can easily solve all the above three points. First of all, users don't need to do anything, when users click "Start Presentation", the computer will automatically turn into Presentation Mode, secondly, when the desktop application is running, unless it is forced to shut down, basically it will not black screen. Finally, the speaker does not need to hold anything in his hands when he is in Slidecontroller, so it does not affect the overall appearance.

6. CONCLUSIONS

Slide controller is a desktop application designed to assist presenters in becoming more confident in their presentations. Implementing "Hands-Free ", allows speakers to use their hand gestures to make their presentations more interesting and to reduce their worries about slide presentations. Users are not limited to a single keyword option, they can create their Transition Keyword to suit the atmosphere of their presentation, and if they don't need it, then Slide controller's Default Keywords will work just as well. Anyone can use this software, and the ultimate goal of Slide controller is to help you deliver a better presentation [15]. Because of the limitations of the current AI voice recognition technology, there is no way to make greater use of Slidecontroller's advantage is to switch slides in one to two seconds or less. To complement this shortcoming, I asa developer will try to train my speech recognition system, and refine it to be sensitive and accurate.

REFERENCES

[1] Pittenger, Khushwant KS, Mary C. Miller, and Joshua Mott. "Using real-world standards to enhance students' presentation skills." Business Communication Quarterly 67.3 (2004): 327-336.

[2] Hayama, Tessai, Hidetsugu Nanba, and Susumu Kunifuji. "Structure extraction from presentation slide information." Pacific Rim International Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2008.

[3] McGehee, Frances. "An experimental study of voice recognition." The Journal of General Psychology 31.1 (1944): 53-65.

[4] Turban, Georg, and Max Mühlhäuser. "A uniform way to handle any slide-based presentation: the universal presentation controller." International Conference on Multimedia Modeling. Springer, Berlin, Heidelberg, 2006.

[5] Hornaday, John A., and John Aboud. "Characteristics of successful entrepreneurs." Personnel psychology (1971).

[6] Haber, Richard J., and Lorelei A. Lingard. "Learning oral presentation skills." Journal of general internal medicine 16.5 (2001): 308-314.

[7] Steinman, Ralph M. "Dendritic cells and the control of immunity: enhancing the efficiency of antigen presentation." The Mount Sinai journal of medicine, New York 68.3 (2001): 160-166.

[8] Extance, Andy. "How AI technology can tame the scientific literature." Nature 561.7722 (2018): 273-275.

[9] Zhang, Li, et al. "Accelword: Energy efficient hotword detection through accelerometer." Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. 2015.

[10] Ran, Duan, Wang Yingli, and Qin Haoxin. "Artificial intelligence speech recognition model for correcting spoken English teaching." Journal of Intelligent & Fuzzy Systems 40.2 (2021): 3513-3524.

[11] Huang, Yiteng, et al. "Multi-microphone adaptive noise cancellation for robust hotword detection." (2019).

[12] Grønli, Tor-Morten, et al. "Mobile application platform heterogeneity: Android vs Windows Phone vs iOS vs Firefox OS." 2014 IEEE 28th International Conference on Advanced Information Networking and Applications. IEEE, 2014.

[13] McDonald, Robert S., and Paul A. Wilks. "JCAMP-DX: A standard form for exchange of infrared spectra in computer readable form." Applied Spectroscopy 42.1 (1988): 151-162.

[14] Haque, Md, Abu Raihan, and Mohd Khalidi. "BTpower: An Application for Remote Controlling PowerPoint Presentation Through Smartphone." Advances in Computer and Computational Sciences. Springer, Singapore, 2017. 13-20.

[15] Alley, Michael, et al. "How the design of headlines in presentation slides affects audience retention." Technical communication 53.2 (2006): 225-234.

A BELIEF REVISION MECHANISM WITH TRUST REASONING BASED ON EXTENDED RECIPROCAL LOGIC FOR MULTI-AGENT SYSTEMS

Sameera Basit and Yuichi Goto

Department of Information and Computer Science, Saitama University,
338-8570 Saitama, Japan

ABSTRACT

When an agent receives messages from other agents, it does belief revision. A belief revision includes, i) a trust reasoning process, i.e., it obtains new belief related to the messages, and deduces implicitly unknown beliefs from the obtained belief; ii) in the case of contradiction in the belief set, it resolves the contradiction. So, trust reasoning, and belief revision must be included in the decision-making process of an intelligent agent in multi-agent systems. Although a belief revision mechanism with trust reasoning is demanded to construct multi-agent systems, there is no such belief revision mechanism. We, therefore, present a belief revision mechanism with trust reasoning based on extended reciprocal logic for multi-agent systems.

KEYWORDS

Multi-agent Systems, Trust Relationship, Trust Reasoning, Strong Relevant Logics, Belief Revision

1. INTRODUCTION

A trust relationship is one of the important reciprocal relationships in our society and cyberspace. Many reciprocal relationships must concern two parties [1]. Especially, the trust relationship is the basis of communications among agents (human to human, human to system, and system to system), and the basis of the decision-making of the agents.

Trust reasoning must be included in the decision-making process of an agent with reasoning capability, an intelligent agent for short, in multi-agent systems. Trust reasoning is a process to draw propositions from already known propositions using the degree of trust of an agent or a received message. A belief of an agent is a proposition that the agent believes, i.e., observed facts, already given theories and assumptions. Any agent in multi-agent systems can extend its belief set by receiving messages from other agents and observing its external environment or own internal status. Especially, an intelligent agent deduces implicitly included propositions from its belief set. After that, the agent decides the next actions according to its current belief set. An intelligent agent in an open system should be able to change the way it handles messages from other agents depending on the degree of trust of the agents because not all agents in the system can be trusted. Thus, an intelligent agent should be able to do trust reasoning for its decision-making.

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 161-173, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130413

Belief revision must also be included in the decision-making process of an agent in multi-agent systems. Belief revision is a process of solving a contradiction in a target belief set to keep the belief set consistent. A belief set is consistent if and only if the set does not include both a proposition and its negation. In an open multi-agent system in the real world, the belief set of an agent is not always consistent, because a given assumption and an observed fact, or a previously observed fact and the current observed fact are sometimes explicitly or implicitly contradicted. Thus, an agent should be able to do belief revision. Moreover, in general, a trust relationship is not an eternal relationship. Although an agent is trusted at a point in time, the agent will not be trusted at another point in time. Changing trust relationships among agents, an agent updates its belief set by belief revision.

Although a belief revision mechanism with trust reasoning is demanded to construct multi-agent systems, there is no such belief revision mechanism. On one hand, the best-known work on modeling belief revision is the so-called Alchourrón, Gärdenfors, and Makinson's (AGM) theory or AGM model[2,3,4]. The AGM model is not suitable for the belief revision mechanism with trust reasoning because the AGM model adopts classical mathematical logic [5]. Classical mathematical logic is a suitable logic system underlying proving but not reasoning [5]. On the other hand, a well-known belief revision mechanism is the so-called truth maintenance systems, belief revision systems, or reason maintenance systems [6]. Essentially, the concept of truth maintenance systems is independent of a specific logic system. However, there is no truth maintenance system based on a logic system underlying trust reasoning.

This paper presents a belief revision mechanism with trust reasoning based on extended reciprocal logic for multi-agent systems. The belief revision mechanism is a Doyle's-style approach (truth maintenance system approach) to deal with the inconsistency in an agent's belief set. The mechanism consists of two parts. First, trust reasoning based on extended reciprocal logic is applied to the deduction process. Extended reciprocal logic is a candidate fora suitable logic system underlying trust reasoning. The second part deals with the belief revision of each agent in multi-agent systems. The proposed mechanism uses the concept of a derivation path. A derivation path can be viewed as a representation of a belief set that is gradually developed and modified as a result of changes in trust relationships with other agents. If a contradiction occurs in the belief set, a revision process is triggered which allows forward and backtracking within the derivation path to track beliefs that cause inconsistency in the agent's belief set.

The rest of the paper is organized as follows: Section 2 shows extended reciprocal logic as a suitable logic system underlying trust reasoning. Section 3 describes a belief revision mechanism with trust reasoning based on extended reciprocal logic. Section 4 illustrates the application of the belief revision mechanism. Some concluding remarks are given in section 5.

2. EXTENDED RECIPROCAL LOGIC

A logic system underlying trust reasoning should be able to deal with various trust properties. A trust relationship consists of a trustor, a trustee, and the trust property, indicating that the trustor believes that the trustee satisfies the trust property [7]. In the context of trust, not all the information from the other agent can be taken as a true message, i.e., "an agent α trusts another agent β with respect to a certain property" means that "α believes that β satisfies this property." Demolombe [8] defined several trust properties. His definitions are as follows.

- *Sincerity*: An agent α trusts in the sincerity of an agent β if f if β informs α about a proposition p then β believes p.
- *Validity*: An agent α trusts in the validity of an agent β if f if β informs α about a proposition p then p is the case.

- *Completeness*: An agent α trusts in the completeness of an agent βiffif p is the casethen β informs α about p.
- *Cooperativity*: An agent α trusts in the cooperativity of an agent βiffif β believes p then β informs *alpha* about p.
- *Credibility*: An agent α trusts in the credibility of an agent β iffif β believes p then p is the case.
- *Vigilance*: An agent α trusts in the vigilance of an agent β iffif p is the case then β believes p.

Trust reasoning is a process to draw propositions from already known propositions using the degree of trust of an agent or a received message. Thus, a logic system underlying trust reasoning should be able to deal with such trust properties.

A logic system underlying trust reasoning should be suitable for forward reasoning. Classical mathematical logic and its various conservative extensions are not suitable for logic systems underlying reasoning because they have paradoxes of implication [9, 10]. Strong relevant logic has rejected those paradoxes of implication and is considered the universal basis of various applied logic for knowledge representation and reasoning [5]. Thus, strong relevant logic and its conservative extensions are candidates for logic systems underlying reasoning. Reciprocal logic [1] is one of the conservative extensions of strong relevant logic to deal with various reciprocal relationships, including trust relationships. However, the reciprocal logic cannot deal with the trust properties [11, 12].

Therefore, a logic system underlying trust reasoning, named extended reciprocal logic, was proposed [11, 12]. Extended reciprocal logic, ERL for short, is an extension of reciprocal logic by introducing trust properties, i.e., sincerity, validity, completeness, cooperativity, credibility, and vigilance, to the reciprocal logic. The extended reciprocal logic is a hopeful candidate for a logic system underlying trust reasoning.

ERL consists of several predicates, two modal operators, and several axioms added to the reciprocal logic. Since ERL is one of the conservative extensions of strong relevant logic, ERL adopts all logical theorems of strong relevant logic. ERL also adopts all logical theorems of reciprocal logic. Below are the modal operators, predicates for representing messages, axioms, and inference rules of ERL.

Modal Operators are as follows.

- $Bel_i(p)$: agent I believes that a proposition p is true.
- $Inf_{i,j}(p)$: agent I has informed agent j about p.

ERL provides a predicate $TR(pe_1,pe_2,PROP)$ where pe_1 and pe_2 are agents, and $PROP$ is an individual constant that represents trust properties: sincerity, validity, completeness, cooperativity, credibility, and vigilance in extended reciprocal logic. For example, $TR(pe_1,pe_2,sincerity)$ means "pe_1 trusts pe_2 in sincerity", $TR(pe_1,pe_2,credibility)$ means "pe_1 trusts pe_2 in credibility", $TR(pe_1,pe_2,completeness)$ means "pe_1 trusts pe_2 in completeness", and in the same way, we can define a predicate for other trust properties as well. Additionally, $TR(pe_1,pe_2,all)$ means "pe_1 trusts pe_2 in all trust properties".
Axioms are as follows.

*ERcL*1: $\forall i \forall j(TR(i,j,sincerity) \Rightarrow (Inf_{j,i}(A) \Rightarrow Bel_j(A)))$
*ERcL*2: $\forall i \forall j(TR(i,j,validity) \Rightarrow (Inf_{j,i}(A) \Rightarrow A))$
*ERcL*3: $\forall i \forall j(TR(i,j,vigilance) \Rightarrow (A \Rightarrow Bel_j(A)))$

*ERcL*4: $\forall i \forall j(TR(i,j,credibility) \Rightarrow (Bel_j(A) \Rightarrow A))$
*ERcL*5: $\forall i \forall j(TR(i,j,cooperativity) \Rightarrow (Bel_j(A) \Rightarrow Inf_{j,i}(A)))$
*ERcL*6: $\forall i \forall j(TR(i,j,completeness) \Rightarrow (A \Rightarrow Inf_{j,i}(A)))$
BEL: $\forall i(Bel_i(A \Rightarrow B) \Rightarrow (Bel_i(A) \Rightarrow Bel_i(B)))$
ERL has three inference rules: modus ponens $\Rightarrow E$,adjunction $\wedge I$, and necessitation *Bel* − *Nec*. The two of three rules come from strong relevant logic. The *Bel* − *Nec*is introduced to the reciprocal logic.

$\Rightarrow E$: "from A and $A \Rightarrow B$ to infer B" (Modus Ponens)
$\wedge I$: "from A and B infer $A \wedge B$" (Adjunction)
Bel − *Nec*: "if A is a logical formula, then so is $Bel_i(A)$" (Necessitation)
Conclusively, ERL is $RcL \cup \{ERcL1,...,ERcL6,BEL\}$ where *RcL* is all axioms of the reciprocal logic. Trust reasoning based on ERL is deductive reasoning from given logical formulas and all logical theorems of ERL.

3. BELIEF REVISION MECHANISM WITH TRUST REASONING BASED ON EXTENDED RECIPROCAL LOGIC

An agent in a multi-agent system has a set of beliefs as observed facts, previously given theories, and hypotheses. Using a set of beliefs, the agent calculates trust relationships between other agents by using trust reasoning within the domain to determine which agent should be trusted by the agent. When the agent receives messages from other agents, it does belief revision. Each time an agent in a domain receives a message from another agent, it undergoes a series of steps, as depicted in figure 1. The belief revision mechanism is comprised of two stages, as of the first stage it undergoes a trust reasoning process, i.e., it obtains new beliefs related to the messages, and deduces implicitly unknown beliefs from the obtained beliefs. These beliefs become part of the agent's belief set. In the second stage, if the deduced beliefs contradict pre-existing beliefs in the agent's belief set, it resolves the contradiction to maintain consistency.

In our belief revision mechanism, if a contradictory belief is entered into the belief set, a revision procedure is initiated to work backward through the path following the belief contained in the label, seeking to determine which belief may have contributed to the contradiction. In order to eliminate the contradiction, some of the existing beliefs are removed from the set of beliefs, and use the labels once again to remove all deductions that originated from these beliefs from the set of current beliefs. Although this process may result in some complexity issues, it is nevertheless theoretically feasible. Details of each sub-process of the belief revision mechanism are discussed in the following sections.

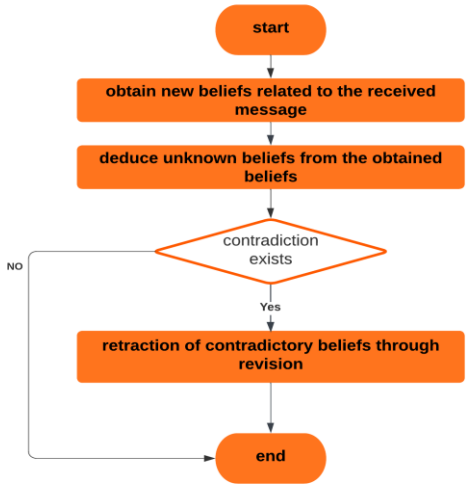


Figure 1.Belief revision process with trust reasoning

3.1. Creation of logical formulas to obtain new beliefs related to the received message

Upon receiving messages from other agents within a domain, new beliefs are obtained by generating logical formulas. To generate a logical formula that indicates that an agent has informed another agent about a message, e.g., "*m is valid*" is a message informed by agent b to agent a, and as a predicate, it will be represented as *is Valid(m)*. Then its related logical formula will be generated as $Inf_{b,a}$ *(is Valid(m))*. From these generated logical formulas new beliefs, e.g., Bel_b *(is Valid(m))* is obtained.

3.2. Deduction of unknown beliefs from the obtained beliefs

Through trust reasoning using axioms, and inference rules from the ERL. This deduced implicit unknown beliefs from the obtained beliefs, and this deduced belief becomes the part of agent's belief set. Each agent maintains a belief set as a derivation path. Deduced beliefs are entered into the derivation path. As a result of the deduction process, an agent gradually adds or modifies its beliefs. As new beliefs are added to the belief set at each time instance, the derivation path evolves over time. Additionally, the derivation path identifies which inference rule was utilized, as well as which beliefs were used as premises or sources using the labeled formula concept.

A deduced belief in a derivation path is labeled with the time stamp, i.e., an integer indicating the instance at which this occurred. The time stamp serves as an index indicating the logical formula position in the belief set. Since these deduced beliefs are derived from premises using inference rules. These labels contain a record of which inference rule was used, as well as which beliefs were used as premises, or sources. This way the agent knows all the logical consequences of each logical formula in his belief set. A label is defined as an ordered 4-tuple (index, from, to, status) [13], where :

1. index is a non-negative integer, the index, representing the position of the deduced belief in the belief set.
2. from-list contains information about premises, and inference rules used to derive thededuced belief.

3. to-list contains an index of all deduced beliefs where the given deduced belief servesas a premise.
4. status, using values *on* and *off*, indicates that only beliefs with status *on* can be used as premises in the deduction process. Whenever a deduced belief is first entered into the belief set, it is assigned status *on*.

3.3. Retraction of Contradictory Belief

Trust reasoning deduces beliefs that sometimes contradict pre-existing beliefs in the agent's belief set. Upon contradiction, a revision procedure is triggered, which disbelieves previously held beliefs, thus retracting the belief set by the contradictory belief. Usually, beliefs can be obtained as a message received from other agent in a domain, or it can be derived from the trust reasoning process. The procedure has three steps:

1. By backtracking through the belief set, starting with the from-list in the label of the contradictory belief, identify the beliefs that were involved in the derivation of the contradictory belief causing inconsistency in the belief set.
2. Change the status of involved beliefs to *off*, as many as necessary to invalidate the derivation of the given contradictory belief. The decision as to which status to turn *off* can be decided by retracting the one that is least believed generally identified by epistemic entrenchment value [3]. In cases where all the involved beliefs are equally believed, a random choice can be made. In some systems, this retraction process may be automated, and in others, it may be human-assisted [15].
3. Forward chains using the to-lists, identify all beliefs whose derivations were based on the retracted belief, and put their status to *off* as well.

This retraction of beliefs will include those beliefs that cause the agent's belief set to be inconsistent. Changing a belief's status from *on* to *off* occurs whenever a contradiction occurs. The objective of the revision procedure is to remove such contradictory beliefs from the agent's belief set.

The following sections will discuss the application of the belief revision mechanism in two case studies, a scenario about public key infrastructure, and a scenario about a spy novel.

4. APPLICATION OF THE BELIEF REVISION MECHANISM IN PUBLIC KEY INFRASTRUCTURE

As an example, we demonstrated the application of the belief revision mechanism in public key infrastructure PKI. When a change in trust relationships occurs between agents, it affects the trust reasoning process, and as a result, it deduces different results from trust reasoning. Following is the public key infrastructure scenario depicting trust relationships, and the exchange of messages between agents.

4.1. Public key infrastructure PKI scenario

In the PKI scenario, agents e_1, e_2, and e_3 exchange messages as certificates among themselves. Agent e_1 is informed about certificate $c_1$by the parent of the agent. We consider that every agent trusts its parent agent in its validity. Furthermore, agents e_2 and e_3 inform agent e_1 about certificates c_2 and c_3 respectively. Agent $e1$ doesn't believe the certificates c_2 and $c_3$but wishes to use them. Therefore, based on the trust relationships between agents, messages such as certificates can be reasoned out as beliefs through trust reasoning. Moreover, taking into

consideration that agent $e_4$informs that c_1 is not valid, here if the deduced belief through the trust reasoning process contradicts the existing beliefs of agent e_1 belief set revision process will be invoked.

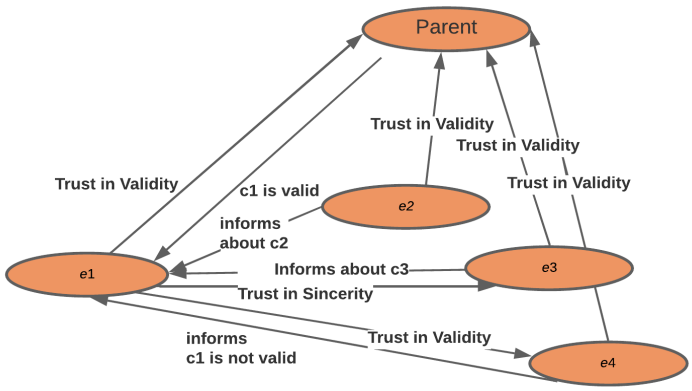


Figure 2. Public key infrastructure scenario

4.2. Formalization

To formalize the above scenario, we defined the following constants, functions, and predicates.

Individual variables:

- e: an agent
- c, c': certifications

Individual constants:

- e_1,e_2, e_3, e_4: agents
- c_1, c_2, c_3, c_4: certifications
- *today*: date of today

Functions:

- $I(c)$: Issuer of certification c.
- $S(c)$: Subject of certification c.
- $PK(c)$: Public key of c.
- $SK(c)$: Share key of c.
- $DS(c)$: Start date of c.
- $DE(c)$: End date of c.
- $Sig(c)$: Signature of c.
- $parent(e)$: The parent of agent e.

Predicates:

- $inCRL(c)$: c is in the certification revocation list.
- $isValid(x)$: x is valid.
- $isSigned(x,k)$: x is message signed by key k.
- $x = y$: x is equal to y.

- $x \leq y$: x is equal to or less than y.
- $x < y$: x is less than y.

Empirical theories of PKI

We can assume the following empirical theories.
PKI1: $\forall e(TR(e,parent(e),validity))$
(Any agent trusts its parent agent in validity.)
PKI2: $\forall c(\exists c'((isValid(c'))) \wedge (I(c) = S(c')) \wedge (isSigned(c,PK(c')))) \Rightarrow isValid(Sig(c)))$
PKI3: $\forall c((isValid(Sig(c)) \wedge (DS(c) \leq today) \wedge (today < DE(c)) \wedge \neg inCRL(c)) \Rightarrow isValid(c))$
(PKI2 and PKI3 allow to verify the signature, and certificate itself on the basis of another certificate whose validity has been proven.)

Logical theories

We can assume the following logical formulas.
P1-1: $I(c_2) = S(c_1)$, P1-2: $I(c_3) = S(c_1)$
(These observed facts are used as premises in our reasoning process and it is true in this scenario only.)
P2-1: $isSigned(c_2,PK(c_1))$
P2-2:$isSigned(c_3,PK(c_1))$
(A certificate $c_2$orc_3 is signed by the subject of certificate c_1 with the private key corresponding to the public key of c_1.)
P3-1: $Inf_{parent(e1),e1}(isValid(c_1))$,
P3-2: $Inf_{e3,e1}(isValid(c_3))$
(The parent agent of e_1 has informed e_1 about "certificate c_1 is valid".)
P3-3: $Inf_{e4,e1}(\neg isValid(c_1))$
P4: $TR(e_1,e_3,sincerity)$ (*assumption*)
P4-1: TR $(e_1,e_4, validity)$ (*assumption*)
P5-1: $DS(c_2) \leq today$, P5-2: $DS(c_3) \leq today$(*assumption*)
P6-1: $today < DS(c_2)$, P6-2: $today < DS(c_3)$ (*assumption*)
P7-1: $\neg inCRL(c_2)$, P7-2:$\neg inCRL(c_3)$ (*assumption*)

4.3. Trust reasoning process

Case 1: Agent$e_1$received certificate $c_1$as a message from its parent.

1. $Inf_{parent(e1),e1}(isValid(c_1)) \Rightarrow isValid(c_1)$ [from PKI1, ERcL2 with $\Rightarrow E$]
2. $isValid(c_1)$ [from P3-1, 2]
3. $Bel_{e1}(isValid(c_1))$ [from 2 with *Bel* − *Nec*]

Case 2: Agent$e_1$received certificate $c_2$as message from agente_2

4. $(I(c_2) = S(c_1)) \wedge isSigned(c_2,PK(c_1))$ [from P1-1 and P2-1 with $\wedge I$]
5. $Bel_{e1}((I(c_2) = S(c_1)) \wedge isSigned(c_2,PK(c_1)))$ [from 4 with *Bel* − *Nec*]
6. $(isValid(c_1) \wedge (I(c_2) = S(c_1)) \wedge (isSigned(c_2,PK(c_1))) \Rightarrow isValid(Sig(c_2))$ [Replaced c with c_2 and c' with c_1 in PKI2]
7. $Bel_{e1}(isValid(c_1) \wedge (I(c_2) = S(c_1)) \wedge (isSigned(c_2,PK(c_1)))) \Rightarrow isValid(Sig(c_2)))$ [from 6 with *Bel* − *Nec*]
8. $Bel_{e1}(isValid(c_1) \wedge (I(c_2) = S(c_1)) \wedge (isSigned(c_2,PK(c_1)))) \Rightarrow Bel_{e1}(isValid(Sig(c_2)))$ [from *BEL* and 7 with $\Rightarrow E$]

9. $Bel_{e1}(isValid(Sig(c_2)))$ [from 5 and 8 with $\Rightarrow E$]
10. $Bel_{e1}(DS(c_2) \leq today)$, $Bel_{e1}(today < DS(c_2))$, $Bel_{e1}(\neg inCRL(c_2))$ [from each of P5-1, P6-1, and P7-1 with *Bel − Nec*]
11. $Bel_{e1}(isValid(Sig(c_2)) \wedge (DS(c_2) \leq today) \wedge (today < DE(c_2)) \wedge \neg inCRL(c_2))$ [from 10 with $\wedge I$]
12. $isValid(Sig(c_2)) \wedge (DS(c_2) \leq today) \wedge (today < DE(c_2)) \wedge \neg inCRL(c_2)) \Rightarrow isValid(c_2)$ [Replaced *c* with c_2 in PKI3]
13. $Bel_{e1}(isValid(Sig(c_2)) \wedge (DS(c_2) \leq today) \wedge (today < DE(c_2)) \wedge \neg inCRL(c_2)) \Rightarrow isValid(c_2)$ [from 12 with *Bel − Nec*]
14. $Bel_{e1}(isValid(Sig(c_2)) \wedge (DS(c_2) \leq today) \wedge (today < DE(c_2)) \wedge \neg inCRL(c_2)) \Rightarrow Bel_{e1}(isValid(c_2))$ [from *BEL* and 13 with $\Rightarrow E$]
15. $Bel_{e1}(isValid(c_2))$ [from 11 and 14 with $\Rightarrow E$]

In cases 1 and 2, beliefs $Bel_{e1}(isValid(c_1))$ and $Bel_{e1}(isValid(c_2))$ are deduced from the trust reasoning process, and these deduced beliefs will be entered into the agent's belief set with their labels, i.e. labels of beliefs $Bel_{e1}(isValid(c_1))$ and $Bel_{e1}(isValid(c_2))$ will be (3, (2, *Bel − Nec*), (7, 8), On), and (15, (11, 14, =>E), {}, On) respectively.

Case 3: Agent e_1 received certificate c_3 as a message from agent e_3.

16. $isValid(c_1) \wedge (I(c_3) = S(c_1)) \wedge isSigned(c_3, PK(c_1))$ [from 2, P1-2, and P2-2 with $\wedge I$]
17. $\exists c'((isValid(c')) \wedge (I(c_3) = S(c')) \wedge (isSigned(c_3, PK(c'))) \Rightarrow isValid(Sig(c_3))$ [Substitute c_3 for *c* in PKI2]
18. $isValid(Sig(c_3))$ [from 16 and 17 with $\Rightarrow E$]
19. $isValid(Sig(c_3)) \wedge (DS(c_3) \leq today) \wedge (today < DE(c_3)) \wedge \neg inCRL(c_3)$ [from 18 and P5-2, P6-2, and P7-1 with $\wedge I$]
20. $isValid(Sig(c_3)) \wedge (DS(c_3) \leq today) \wedge (today < DE(c_3)) \wedge \neg inCRL(c_3) \Rightarrow isValid(c_3)$ [Substitute c_3 for *c* in PKI3]
21. $isValid(c_3)$ [Deduced from 19 and 20 with $\Rightarrow E$]
22. $Inf_{e3,e1}(A) \Rightarrow Bel_{e3}(A)$ [from P3-2 and ERcL1 with $\Rightarrow E$]
23. $Bel_{e1}(isValid(c_3)$ [from P4 and 22 with $\Rightarrow E$]

In case 3, $Bel_{e1}(isValid(c_3))$ is deduced from the trust reasoning process, and deduced belief will be entered into the agent's belief set with its respective label (23, (P4, 21, =>E), {}, On).

Case 4: Agent e_1 received a message about the certificate c_1 from agent e_4

24. $Inf_{e4,e1}(\neg isValid(c_1)) \Rightarrow \neg isValid(c_1)$ [from P4-1, ERcL2 with $\Rightarrow E$]
25. $\neg isValid(c_1)$ [from P3-3, 25]
26. $Bel_{e1}(\neg isValid(c_1))$ [from 25 with *Bel − Nec*]

In case 4, $Bel_{e1}(\neg isValid(c_1))$ is deduced, and deduced belief will be entered into the agent's belief set with its respective label (26, (25, =>E), {}, On).

4.4. Revision process under the belief revision mechanism

Belief set of agent e_1 represented as BS_{e1}={}. Initially, the belief set will be empty as $BS_{e1} = \phi$. Based on the current scope of study beliefs can be obtained in two ways, i) A belief can be received as a message from other agents in the domain; ii) A belief can be derived as a deduced belief from the trust reasoning process, i.e., change in trust relationship deduces different

reasoning results. So, until four beliefs are part of the agent belief set. Currently,agente_1 belief set has Bel_{e1}={Bel_{e1}(*isValid*(c_1)),Bel_{e1}(*isValid*(c_2)),Bel_{e1}(*isValid*(c_3), Bel_{e1}(¬*isValid*(c_1))}.

Beliefs are retained in the agent's belief set with their labels which helps to maintain the derivation path.Entries of other beliefs are handled in a similar manner. Now the belief set of agent e1 consists of two contradictory beliefs along with their labels. i.e., Bel_{e1}(*isValid*(c_1)) and Bel_{e1}(¬*isValid*(c_1)). So, the revision process in section 3.0.3 will be triggered to retract the contradictory belief. If belief Bel_{e1}(*isValid*(c_1)) is selected as discussed in point 2 of section 3.0.3, then the revision procedure forward chains through to-lists, changing the status of deduced belief at 7, and 8 from *on* to *off*. To this point, beliefs Bel_{e1}(*isValid*(c_1)),Bel_{e1}(*isValid*(c_2)) will have their statuses *off*, leaving BS_{e1}={Bel_{e1}(*isValid*(c_3), Bel_{e1}(¬*isValid*(c_1))} in belief set of agente_1. Using this method, agents would retain their beliefs, but their status would be set to*off*. As a result, it will be possible to trace the beliefs, but at the same time prevent the agent from re-acquiring them, therefore making belief revisions a practical, and useful process.

5. APPLICATION OF THE BELIEF REVISION MECHANISM IN SPY NOVEL

5.1. Spy novel scenario

We consider another scenario from [8] in which multiple agents exchange messages with each other as an information source.

We consider three agents a_1, b_1, and c_1 who are interested in exchanging information about the two facts "there is a spy in the train T", denoted by p_1, and "the train T has arrived at the railway station", denoted by q. In this situation agent a_1 trusts b_1 in regard to his validity for p_1, and in regard to his sincerity for q_1, and a_1 trusts c_1 in regard to his completeness for q_1. a_1 trust may be supported, for instance, by the fact that *b* belongs to some intelligence service, and c_1 is an employee of the railway station who stands on the platform where the train is supposed to arrive. In this situation, b_1 has informed a_1 of information p_1, and he has also informed q_1, and c_1 has not informed a_1 of information q_1.The formalization of the above scenario is as follows:

5.2. Formalization

Individual variables:

- *agents: a,b,c*
- *facts: p, q*

Individual constants:

- *agents: a1,b1,c1*
- *facts: p1, q1*

Predicates:

- *isFact(x): x is a fact.*

Empirical and logical theories

We can assume the following theories.
IS1: *TR*(a_1,b_1,*validity*)(Agent a_1 trusts b_1 in his validity)

IS2: $TR(a_1,b_1,sincerity)$(Agent a_1 trusts b_1 in his sincerity)
IS3: $TR(a_1,c_1,completeness)$(Agent a_1 trusts c_1 completeness)
IS3-1: $TR(a_1,c_1,sincerity)$ *(Agent* a_1 trusts c_1 sincerity)
IS4: $Inf_{b1,a1}(isFact(p_1))$(b_1 has informed to a_1 about $isFact(p_1)$)
IS5: $\neg Inf_{c1,a1}(isFact(q_1))$($c_1$ has not informed to a_1 about $isFact(q_1)$)
IS6: $Inf_{c1,a1}(\neg isFact(q_1))$($c_1$ has informed to a_1 about $\neg isFact(q_1)$)
IS7: $\neg Inf_{b1,a1}(isFact(q_1))$($b_1$ has not informed to a_1 about $isFact(q_1)$)

5.3. Trust reasoning process

From the above formalization, empirical and logical theories obtained as logical formulas will be used in the trust reasoning process
.

Case 1: Agent a_1 received information about p_1 as a message from agent b_1.

1. $Inf_{b1,a1}(isFact(p_1)) \Rightarrow isFact(p_1)$ [from *IS*1 and *ERcL*2 with ⇒E]
2. $isFact(p_1)$[from IS4 and 1 with ⇒E]
3. $Bel_{a1}(isFact(p_1))$[from 2with*Bel –Nec*]

After deduction, we have $Bel_{a1}(isFact(p_1))$. The deduced belief will be added to the belief set of agents a_1 with its respective label (3, (2, *Bel – Nec*), {11}, On).

Case 2: Agent a_1 received information about q_1 as a message from agent c_1.

4. $Bel_{a1}(\neg Inf_{c1,a1}(isFact(q_1)))$[from IS5 with*BEL –Nec*]
5. $A \Rightarrow Inf_{c1,a1}(A)$ [from IS3 and ERcL6 with ⇒ E]
6. $isFact(q_1) \Rightarrow Inf_{c1,a1}(isFact(q_1))$ [from 5]
7. $\neg Inf_{c1,a1}(isFact(q_1) \Rightarrow \neg isFact(q_1)$ [*contraposition* of 6]
8. $Bel_{a1}(\neg Inf_{c1,a1}(isFact(q_1) \Rightarrow \neg isFact(q_1))$ [from 7 with *BEL – Nec*]
9. $Bel_{a1}(\neg Inf_{c1,a1}(isFact(q_1)) \Rightarrow Bel_{a1}(\neg isFact(q_1))$ [from 8 with *BEL*]
10. $Bel_{a1}(\neg isFact(q_1))$ [from 4and 9with ⇒ E]
11. $Bel_{a1}(isFact(p_1) \wedge \neg isFact(q_1))$ [from 3and10 with ∧*I*]

After deduction we have $Bel_{a1}(isFact(p_1) \wedge \neg isFact(q_1))$. The deduced belief will be added to the belief set of agents a_1 with its respective label (11, (3, 10, ∧*I*), {}, On).

Case 3: Agent a_1 received information about p_1 as a message from agent $c_1$with a change in a trust relationship.

12. $Inf_{b1,a1}(A) \Rightarrow Bel_{b1}(A)$ [from IS2 and ERcL1 with ⇒*E*]
13. $Inf_{b1,a1}(isFact(p_1)) \Rightarrow Bel_{b1}(isFact(p_1))$ [from 12]
14. $Bel_{b1}(isFact(p_1))$[from IS4 and 13 with ⇒*E*]
15. $Bel_{a1}(Inf_{c1,a1}(\neg isFact(q_1))$ [from IS6with*Bel –Nec*]
16. $Inf_{c1,a1}(A) \Rightarrow Bel_{c1}(A)$ [from IS3-1 and ERcL1 with ⇒*E*]
17. $Inf_{c1,a1}(\neg isFact(p_1)) \Rightarrow Bel_{c1}(\neg isFact(p_1))$ [from 16]
18. $Bel_{c1}(\neg isFact(p_1))$ [from IS6 and 17 with ⇒*E*]
19. $Bel_{a1}(Bel_{c1}(\neg isFact(p_1))$ [from 18with *BEL – Nec*]
20. $Bel_{a1}(Bel_{b1}(isFact(p_1)))$ [from 14 with *BEL – Nec*]
21. $Bel_{a1}(Bel_{b1}(isFact(p_1) \wedge Bel_{c1}(\neg isFact(p_1))$ [from 19 and 20 with ∧*I*]

After the trust reasoning process, $Bel_{a1}(Bel_{b1}(isFact(p_1) \wedge Bel_{c1}(\neg isFact(p_1))$ has been deduced. The deduced result will be added to the belief set of agent$a_1$with its respective label (21, (19, 20, $\wedge I$), {}, On). Change in a trust relationship from completeness to sincerity between agent a_1 trusts $c_1$deduces different reasoning results $Bel_{a1}(isFact(p_1) \wedge \neg isFact(q_1))$, and$Bel_{a1}(Bel_{b1}(isFact(p_1) \wedge Bel_{c1}(\neg isFact(p_1))$respectively. Therefore, it is evident from the deduced results that a change in trust relationships leads to different deduced results.

5.4. Revision process under the belief revision mechanism

Initially, the belief set of agentsa_1 is empty $BS_{a1} = \phi$. After the reasoning process, the belief set of agenta_1 will include deduced beliefs, i.e., BS_{a1}={$Bel_{a1}(isFact(p_1) \wedge \neg isFact(q_1))$,$Bel_{a1}(Bel_{b1}(isFact(p_1) \wedge Bel_{c1}(\neg isFact(p_1))$}.As discussed before, a belief can be obtained as a message from another agent in the domain, or it can be derived through the trust reasoning process. So, in the current scenario, if we consider receiving a belief as a message from other agents, and it contradicts the existing beliefs of the agent's a_1 belief set BS_{e1} then the revision process discussed in section 3.0.3 will be triggered to retract the contradictory belief. If belief $Bel_{a1}(isFact(p_1) \wedge \neg isFact(q_1))$ is selected, then the revision procedure forward chains through to, and from lists, changing the status of belief from *on* to *off*. To this point, the contradictory belief causing inconsistency will have their statuses both subsequent beliefs will have their statuses *off*, leaving$Bel_{a1}(Bel_{b1}(isFact(p_1) \wedge Bel_{c1}(\neg isFact(p_1))$ in the belief set of agent a_1. Using this method, agents would retain their beliefs, but their status would be set to*off*. As a result, it will be possible to trace the beliefs, but at the same time prevent the agent from re-acquiring them. Thus, the resulting belief set is consistent.

6. CONCLUDING REMARKS

In this paper, we presented a belief revision mechanism with trust reasoning based on extended reciprocal logic (ERL) for multi-agent systems. A single mechanism that includes trust reasoning, and belief revision for the decision-making process of an agent in multi-agent systems. Trust reasoning based on ERL is used for the deduction process because extended reciprocal logic is a suitable logic system underlying trust reasoning. As a result, an agent maintains its belief set. If a contradiction occurs in the agent's belief set, a revision process based on Doyle's procedural approach is triggered. Doyle's procedural approach uses the concept of derivation path which allows forward, and backtracking to track beliefs that cause inconsistency in the agent's belief set. Furthermore, we demonstrated the application of the belief revision mechanism in the field of public key infrastructure PKI. A unique feature of the belief revision mechanism is that it is based on extended reciprocal logic, which makes it a general mechanism. As part of future work, we will demonstrate the application of the belief revision mechanism in other areas as well.

REFERENCES

[1] J. Cheng, "Reciprocal Logic: Logics for Specifying, Verifying, and Reasoning About Reciprocal Relationships," Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, vol. 3682, pp. 437–445, 2005. doi:10.1007/11552451_58

[2] F. Lévy, "A Survey of Belief Revision and Updating in Classical Logic," International Journal ofIntelligent Systems, vol. 9, no. 1, pp. 29–59, 1994. doi:10.1002/int.4550090104

[3] P. Gärdenfors and H. Rott, "Belief Revision," in Handbook of Logic in Artificial Intelligence and LogicProgramming, vol. 4, D. M. Gabbay, C. J. Hogger, and J. A. Robinson, Eds. Oxford University Press, pp. 35–132, 1995.

[4] L. Sombé, "A Glance at Revision and Updating in Knowledge Bases," International Journal of Intelligent Systems, vol. 9, no. 1, pp. 1–27, 1994, doi:10.1002/int.4550090103

[5] J. Cheng, "A Strong Relevant Logic Model of Epistemic Processes in Scientific Discovery," in Information Modelling and Knowledge Bases XI, E. Kawaguchi, H. Kangassalo, H. Jaakkola, and I. A. Hamid, Eds. Frontiers in Artificial Intelligence and Applications, vol. 61, IOS Press, pp. 136–159, 2000
[6] J. Doyle, "A Truth Maintenance System," Artificial Intelligence, vol. 12, no. 3, pp. 231–272, 1979. doi:10.1016/0004-3702(79)90008-0
[7] W. Zhao, V. Varadharajan, and G. Bryan, "Modelling Trust Relationships in Distributed Environments," Trust and Privacy in Digital Business, Lecture Notes in Computer Science, vol. 3184, pp. 40–49, 2004. doi:10.1007/978-3-540-30079-3_5
[8] R. Demolombe, "Reasoning About Trust: A Formal Logical Framework," Trust Management, LectureNotes in Computer Science, vol. 2995, pp. 291–303, 2004. doi:10.1007/978-3-540-24747-0_22
[9] R. Anderson and N. D. Belnap Jr., "Entailment: the Logic of Relevance and Necessity," vol. 1,Princeton University Press, 1975.
[10] R. Anderson, N. D. Belnap Jr., and J. M. Dunn "Entailment: the Logic of Relevance and Necessity,"vol. 2, Princeton University Press, 1992.
[11] S. Basit and Y. Goto, "An Extension of Reciprocal Logics for Trust Reasoning," Intelligent Information and Database Systems, Lecture Notes in Computer Science, vol. 12034, pp. 65–75, 2020, doi:10.1007/ 978-3-030-42058-1_6
[12] S. Basit and Y. Goto, "An Extension of Reciprocal Logic for Trust Reasoning: A Case Study in PKI,"Intelligent Information and Database Systems, Lecture Notes in Computer Science, vol. 13757, pp. 496–506, 2022. doi:10.1007/978-3-031-21743-2_39
[13] S. Ustymenko and D. G. Schwartz, "Dynamic Agent-oriented Reasoning about Belief and Trust,"Multiagent and Grid Systems, vol. 4, no. 3, pp. 335–346, 2008, doi:10.3233/mgs-2008-4307
[14] D. G. Schwartz, "Nonmonotonic Reasoning as a Temporal Activity," Proceedings of the 15th International Workshop on Non-Monotonic Reasoning, 2014. doi:10.48550/arXiv.1404.7173

AUTHORS

Sameera Basit received her bachelor's degree in Information Technology from Punjab University College of Information Technology, and her master's degree in Computer Science from the Lahore College for Women, University. She is currently a doctoral course student at Saitama University, Japan. Her research interests include intelligent agents, trust in Multi-Agent Systems, trust reasoning, agent decision-making, logic, and software engineering.

Yuichi Goto received the degree of Bachelor of Engineering in computer science, the degree of Master of Engineering in computer science, and the degree of Doctor of Engineering in computer science from Saitama University in 2001, 2003, and 2005, respectively. His current research interests include relevant reasoning and its applications, automated theorem finding, epistemic programming, anticipatory reasoning reacting systems, and Web services.

KNOWLEDGE-ENRICHED MORAL UNDERSTANDING UPON CONTINUAL PRE-TRAINING

Jing Qian[1], Yong Yue[1], Katie Atkinson[2] and Gangmin Li[3]

[1]School of Advanced Technology, Xi'an Jiaotong Liverpool University, China
[2]Department of Computer Science, University of Liverpool, Liverpool, UK
[3]School of Computer Science Technology, University of Bedfordshire, Luton, UK

ABSTRACT

The aim of moral understanding is to comprehend the abstract concepts that hide in a story by seeing through concrete events and vivid characters. To be specific, the story is highly summarized in one sentence without covering any characters in the original story, which requires the machine to behave more intelligently with the abilities of moral perception and commonsense reasoning. The paradigm of "pre-training + fine-tuning" is generally accepted for applying neural language models. In this paper, we suggest adding an intermediate stage to build the flow of "pre-training + continual pre-training + fine-tuning". Continual pre-training refers to further training on task-relevant or domain-specific corpora with the aim of bridging the data distribution gap between pre-training and fine-tuning. Experiments are basing on a new moral story dataset, STORAL-ZH, that composes of 4,209 Chinese story-moral pairs. We collect a moral corpus about Confucius theory to enrich the T5 model with moral knowledge. Furthermore, we leverage a Chinese commonsense knowledge graph to enhance the model with commonsense knowledge. Experimental results demonstrate the effectiveness of our method, compared with several state-of-the-art models including BERT-base, RoBERTa-base and T5-base.

KEYWORDS

Moral Understanding, Continual Pre-training, Knowledge Graph, Commonsense

1. INTRODUCTION

Morality is one of the most complicated topic about humanity [1]. It is tied with commonsense, formed upon ethnic culture, and regulated by rules and laws. Fable stories are must-read books for children, from which they learn morals and ethics to distinguish right from wrong in everyday world. Moral understanding aims to comprehend the abstract concepts that hide in a story by seeing through concrete events and vivid characters, which has become a new challenging task for Natural Language Processing (NLP). Previous works related to story understanding are mainly story ending prediction [2], story completion given constraints (e.g., storylines [3], emotions [4], styles [5], morals [6]). Most of them surround concrete concepts in story itself, whereas our work concentrates on digging out the moral lesson behind it. Table 1 shows one example of moral-story pair from the new dataset STORAL [6].

David C. Wyld et al. (Eds): CCNET, AIMLA, CICS, IOTBS, NLTM, COIT - 2023
pp. 175-185, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130414

Benefit from big data, self-supervised pre-training on an enormous amount of unlabeled corpora from a general domain equips large language models with contextual knowledge and the ability of recognizing n-grams. Originated from Transformer [7], plenty of Pre-trained Language Models (PLMs) have sprung up in succession. They can be roughly categorized in three groups in terms of their architecture, including Transformer encoder (e.g., BERT [8], RoBERTa [9]), Transformer decoder (e.g., GPT2 [10], GPT3 [11]), and the full Transformer encoder-decoder network (e.g., T5 [12], MASS [13]). With dissimilar pre-training strategies, PLMs are suitable to different downstream tasks. For instance, the contextual word representations learned via masked language modeling by RoBERTa are beneficial for natural language understanding, while the strategy of auto-regressive language modeling exploited by GPT2 lays the foundation for natural language generation.

Table 1. An English moral-story pair from STORAL.

Moral	What is evil won is evil lost.
Story	A wolf had stolen a lamb and was carrying it off to his lair to eat it. But his plans were very much changed when he met a lion, who, without making any excuses, took the lamb away from him. The wolf made off to a safe distance, and then said in a much injured tone: “You have no right to take my property like that!” The lion looked back, but as the wolf was too far away to be taught a lesson without too much inconvenience, he said: “Your property? Did you buy it, or did the shepherd make you a gift of it? Pray tell me, how did you get it? ”

The stage of pre-training equips language models with great potential, especially as the number of model parameters and the scale of unlabeled corpora keep growing, which can be proved by the superior performance achieved by prompt learning on a range of benchmark tasks [14]. Prompt learning [15] wraps the input sequence with a template containing masked tokens to handle downstream tasks by imitating the pre-training objectives. By which, the great potential of PLMs is better stimulated. Therefore, continual pre-training on in-domain data (Domain-Adaptive Pre-Training, DAPT) or task-relevant data (Task-Adaptive Pre-Training, TAPT) is a recommended option when the downstream scenarios are of specific domains and no relevant data shows up in the unlabeled corpora.

Other than domain-specific knowledge and task-dependent information that are gained from incremental unlabeled unstructured text, sometimes it is necessary for PLMs to be equipped with the capability of commonsense reasoning. Furthermore, pre-training can be extended to other data of a different structure, such as Knowledge Graph (KG). A typical KG is composed of RDF triples (*h*, *r*, *t*), where *h* and *r* represent head entity and tail entity respectively, *r* represents their relationship. There have been various kinds of KGs, including linguistic [16], encyclopedia [17], commonsense [18], domain-specific [19]. In one popular commonsense KG, ATOMIC [20], triples like (*PersonX applies to jobs*, *xEffect*, *gets hired*), (*PersonX asks PersonY for money*, *xWant*, *to go pay bills*) are telling inferential knowledge about everyday life.

In this paper, we transform the traditional two-stage paradigm of “pre-training + fine-tuning” into three stage by adding an intermediate step of continual pre-training that will be tested by two downstream tasks about moral understanding. We use STORAL-ZH [18], the Chinese part of STORAL [6], as the dataset for target tasks. Furthermore, LongLM-base [21] is selected as our model, that has been pre-trained on 120G Chinese long novels. For TAPT, the language model is further trained on unlabeled STORAL-ZH to equip itself with task-awareness knowledge. For DAPT, we prepare training corpora for two domains including moral culture and commonsense

knowledge. Inspired by [22], we utilize triples of a KG by transforming each triple into one readable textual sequence for continuing to pre-train the language model. To summarize, our contributions are reflected in the following three aspects: (1) Different from the standard paradigm, we choose continual pre-training before fine-tuning to boost model performances on moral understanding. (2) For facilitating the moral perception out of concrete characters and events, we equip our model with the ability of commonsense reasoning by further pre-training on a commonsense KG. (3) We collect a corpus about Chinese traditional moral culture about the Four Books and Five Classics to support domain-adaptive pre-training.

2. RELATED WORK

2.1. Story Understanding

There have been a range of tasks proposed about story understanding and generation, including story ending prediction [2], commonsense story generation [22] and story ending generation with fine-grained sentiment [23]. A variety of attributes are considered for better story understanding, such as storylines [3], emotions [4], styles [5], and morals [6]. Different from that storylines lead the story writting, emotions describe characters ’ states, styles decide the story’s tone, moral understanding aims to discover the implied and abstract theme behind concrete events, that is a more challenging task. [6] firstly proposed moral understanding and generation, and published a new moral story dataset, STORAL.

2.2. Continual Pre-training

Pre-training is definitely the most essential stage for employing language models, that facilitates model initialization and accelerates the parameter convergence on downstream tasks. As the model size grows rapidly, larger-scale unlabeled corpora are required to fully pre-train the model to avoid over-fitting. To bridge the data distribution gap between pre-training and fine-tuning, continual pre-training has been applied and shows to be beneficial for model performance [24, 25]. [26] proposes two concepts about continual pre-training, task-adaptive pre-training (TAPT) and domain-adaptive pre-training (DAPT). The TAPT refers to further pre-training on the unlabeled data of the given task before fine-tuning, which brings consistent improvements [27]. The DAPT requires collecting target domain-relevant corpus, which is probably computationally expensive but still helpful [28].

2.3. Knowledge-Enhanced PLMs

Recently, incorporating knowledge into PLMs is experiencing a surge of interest. Thorough self-supervised pre-training over large-scale corpora provides PLMs with abundant contextual semantics but lacks domain-specific [19, 29], factual [30, 31] or commonsense knowledge [22, 32]. K-BERT [19] explicitly injects triples from domain-specific KG into the input sequence and designs a visible matrix to control the mutual effects among tokens. BERT-MK [29] integrates the graph contextualized knowledge of a medical KG into language models. KEPLER [30] encodes entity descriptions as their embeddings and jointly optimize the knowledge embedding and masked language modeling objectives on the same PLM. ERNIE [31] utilizes the informative entities in KGs to enhance language representation by putting forward a new pre-training objective. KG-BART [32] captures the complex relations of concepts over a commonsense KG for generative commonsense reasoning. [22] conducts incremental pre-training on commonsense knowledge bases to generate more reasonable stories without considering heterogeneous information fusion and sub-graph aggregation, which implicitly and efficiently incorporates commonsense knowledge into GPT-2 [10].

3. METHODOLOGY

This section expatiates the main components of our method, including the PLM, the details about task-adaptive and domain-adaptive pre-training, and the stage of fine-tuning.

3.1. Transformer-based Language Model

The language model adopted in this work is based on the full Transformer architecture [7], where the encoder is fed an input sequence and uses fully-visible masking, the decoder generates the target sequence through causal masking and cross-attention. The text-to-text framework is capable of handling both understanding and generation tasks. One representative encoder-decoder model is T5 [12], it is trained on the Colossal Clean Crawled Corpus (C4) of languages of English, French, Romanian, and German with the best-fit unsupervised pre-training objective of replacing corrupted spans.

A Chinese version of T5, LongLM, is released by [21] after being pre-trained on 120G Chinese novels with two generative tasks, i.e., text infilling [12] and conditional continuation [10]. Inspired by SpanBERT [33], text infilling replaces a few of text spans of input sequence by special tokens with a corruption rate of 15%, while the span lengths are following the Poisson distribution with $\lambda = 3$. Then the target is to output the original text spans replaced by special tokens with the greedy decoding algorithm. The second task, conditional continuation, aims to generate the back half of a text given its front half using top-k sampling [34] with $k = 40$ and a softmax temperature of 0.7 [35]. In this work, we leverage the pre-trained checkpoint of LongLM-base with the number of parameters of 223M on HuggingFace [36].

3.2. Continual Pre-training

3.2.1. Task-Adaptive Pre-Training (TAPT)

To make the pre-trained model more adaptive to downstream tasks, further pre-training on the unlabeled data of the tasks before fine-tuning is worth considering. The advantages of TAPT are reflected in much less computational cost and possible performance boost because the training corpus is far smaller and much more task-relevant. [6] has post-trained the Chinese long-text pre-training model named LongLM [21] on the unlabeled version of STORAL [6], and named it as T5-Post as one compared baseline in the original paper. Table 2 shows an example for the pre-training task of text infilling.

Table 2. An example showing pre-training task of text infilling.

Story	I was sitting in my room and was busy with my usual things. Knowing through the news of social media the carnage of seven civilians, I was afflicted with a heart trouble and great care.
Inputs	I was sitting in my room and was busy with <X>. Knowing through the news of social media <Y>, I was afflicted with a heart trouble and great care.
Targets	<X> my usual things <Y> the carnage of seven civilians <Z>

3.2.2. Domain-Adaptive Pre-Training (DAPT)

Apart from continual pre-training on unlabeled data of downstream tasks, further pre-training on much more unlabeled corpora that are collected from relevant domains is more reasonable. By DAPT, the already powerful PLMs are enriched with additional domain-specific knowledge. As for better moral understanding of fable stories, background domains include moral culture and commonsense knowledge.

Moral Knowledge Confucianism is the mainstream moral culture of China, and the Four Books and Five Classics are its authoritative books, which record in detail the politics, economy, diplomacy, culture and other aspects of the most active period in the development of Chinese ideology, as well as the Confucian philosophy which has influenced Chinese culture for thousands of years. Up to now, the morals and ethics conveyed by the Four Books and Five Classics still regulate, correct and improve the ways we think and behave. The Four Books and Five Classics were written in classical Chinese, we collect the translated version in written vernacular Chinese as the corpus for continual pre-training and named it as **4+5**. Table 3 gives several examples from the Analects out of the Four Books.

Table 3. Three examples from the Analects and translated in Vernacular Chinese and English.

Example 1	不患人之不己知，患不知人也。
Vernacular Chinese	**不要担心别人不了解自己，应该担心的是自己不了解别人。**
English Translation	I am not bothered by the fact that I am unknown. I am bothered when I do not know others.
Example 2	学而不思则罔，思而不学则殆。
Vernacular Chinese	**学习而不思考就会迷惘无所得，思考而不学习就不切于事而疑惑不解。**
English Translation	To study and not think is a waste. To think and not study is dangerous.
Example 3	德不孤，必有邻。
Vernacular Chinese	**品德高尚的人不会孤独，一定有志同道合的人和他做伴。**
English Translation	If you are virtuous, you will not be lonely. You will always have friends.

Table 4. Examples of template-based transformation of KG triples.

Triples	Transformed Sentences
(某人完全放弃某物, **xEffect**, 羞愧地低下头) (*PersonX abandons ____ altogether, **xNeed**, hangs head in shame*)	汤姆完全放弃某物，结果他羞愧地低下头。 Tom abandons something altogether, as a result, he hangs head in shame.
(有人被大学录取了, **xAttr**, 好学的) (*PersonX accepts into college, **xAttr**, studious*)	汤姆被大学录取了，他是好学的。 Tom accepts into college, he is studious.
(某人完成了他的任务, **xIntent**, 赶上最后期限) (*PersonX accomplishes PersonX's task, **xIntent**, to meet a deadline*)	汤姆完成了他的任务，因为他想赶上最后期限。 Tom accomplishes his task, because he wanted to meet a deadline.

Commonsense Knowledge Incorporating commonsense knowledge equips PLMs with the ability of commonsense reasoning in downstream tasks. In this work, we leverage the commonsense knowledge from a structured data, i.e., knowledge graph. ATOMIC [20], one of the most commonly used commonsense knowledge graph, consists of 877K *if-then* triples (*h*, *r*, *t*) in which the head *h* and the tail *t* are two events and the relation *r* describe their *if-then* relationship. For examples, (*PersonX accomplishes PersonY's work*, *xAttr*, *helpful*) means that if *X* accomplishes *Y*'s work, then *X* is helpful, (*PersonX accomplishes PersonY's work*, *oWant*, *to thank PersonX*) tells that if *X* accomplishes *Y*'s work, then *Y* will thank *X*. *xAttr* represents the persona attribute of *X*, *oWant* states others' event. There are three *if-then* types including *If-Event-Then-Mental-State*, *If-Event-Then-Event* and *If-Event-Then-Persona*. Inferential knowledge brought by ATOMIC [20] facilitates language comprehension, especially commonsense relations among concrete events and abstract concepts for moral stories. Our work is based on Chinese, we utilize the translated ATOMIC dataset, ATOMIC-ZH [18] instead. Inspired by [22] and [37], we linearize KG triples into textual sequences through the template-based transformation, as illustrated in Table 4. Different from previous works that explicitly introduced part commonsense knowledge into PLMs, continual pre-training directly on all linearized triples can integrate commonsense knowledge into LongLM [21] implicitly in a more convenient way.

3.3. Fine-Tuning

Following the standard paradigm "pre-training + fine-tuning", we fine-tune our model on two moral understanding tasks after task-adaptive pre-training on unlabeled STORAL-ZH [6] and domain-adaptive pre-training on **4+5** and ATOMIC-ZH [18]. Both tasks are designed by [6], they aim to select the correct moral from several choices given a story, but test the abilities of the PLM from two different aspects. One is concept understanding, the other is preference alignment.

ConcePT understanding (CPT) It requires choosing the correct one from the five candidates of morals for each story, that tests the ability of understanding abstract concepts behind concrete events in the story. Apart from the paired moral of the story, the other four candidates are true negative samples that are selected from the morals of stories about irrelevant topics.

PREFerence alignment (PREF) Simpler than CPT, PREF aims to tell the right moral from the other wrong one. There are only two moral candidates for each story in the constructed task dataset [6]. The incorrect candidate is obtained by replacing one random token in the correct moral with its antonym. As some words do not have antonyms, the training data for PREF is a little smaller than CPT.

To handle both tasks of CPT and PREF, we first concatenate the story and its candidate morals, then insert unique special tokens before the story and each candidate, and feed the sequence into the tested language model. Following the default settings of T5 [12], special tokens are <extra_id_i> where i points out the number order. Inspired by [6], we take the hidden states of corresponding special tokens as the representations of the story and each candidate respectively, afterwards we normalize the dot-product scores between the representations of the story and each candidate to predict the probability distribution over all candidates. We optimize the language model by minimizing the cross-entropy loss.

4. EXPERIMENTS

4.1. Datasets

Corpus for Continual Pre-training We adopt two kinds of corpora of different domains including moral culture and commonsense knowledge for domain-adaptive pre-training. For moral culture, the corpus is composed of vernacular version for the Four Books and Five Classics. The Four Books are Great Learning, Doctrine of the Mean, Analects and Mencius, while the Five Classics are Classic of Poetry, Book of Documents, Book of Rites, I Ching, and Spring and Autumn Annals. We collect the writings in the vernacular of each work from public web resources and integrate them together to get the unlabeled corpus named "**4+5**".

To enrich our model with commonsense knowledge, we transform the triples in ATOMIC-ZH [18] into readable textual sequences using a template-based method [37] for continual pre-training. ATOMIC-ZH [18] is the translated ATOMIC [20] used for Chinese tasks. [18] applies Regular Replacement to alleviate the problems of containing special tokens (i.e., PersonX and PersonY) as well as blank in some triples. To facilitate convenient translation, [18] transform triples into reasonable natural language sentences, then split them into the form of (*h*, *r*, *t*) after being translated via automatic translation system to make up ATOMIC-ZH. The Chinese commonsense knowledge graph provided by [18] is enlarged by other resources, we only select the triples with the nine relations that are mentioned in [20] for our further use.

Corpus for Fine-tuning The corpus for downstream tasks are constructed from STORAL-ZH [6], which composes of 4209 Chinese story-moral pairs. This new dataset is collected by [6] from multiple web pages of moral stories and is cleansed with de-duplication and decoupling. The average number of words and sentences are 322 and 18 for stories, 25 and 1.5 for morals. When applied in the stage of fine-tuning, the labeled data are randomly splitted by 8:1:1 for training/validation/testing set, respectively.

4.2. Compared Baselines

BERT The BERT-architectured model used in our work is the *bert-base-Chinese* register model [8]. It has been pre-trained for Chinese with the pre-training objective of masked language modeling.

RoBERTa The RoBERTa-architectured model used in our work is the *hfl/chinese-roberta-wwm-ext* register model [38]. It is essentially a Chinese pre-trained BERT model with whole word masking.

T5 The T5-architectured model used in our work is the *thu-coai/LongLM-base* register model [21]. It has been pre-trained on 120G Chinese long novels with two pre-training tasks including text infilling [12] and conditional continuation [10].

4.3. Experiment Settings

Our experiments are basing on LongLM-base [21], a Chinese pre-trained T5 model. All language models are implemented on the codes and pre-trained checkpoints from HuggingFace [36]. The model configurations are following their respective base version. As for the hyper-parameters for all models, we set the batch size to 16, the maximum sequence length to 1,024, and the learning rate to 3e-5. As for tokenization, a sentencepiece vocabulary of 32,000 wordpieces [39] is applied. We use accuracy as the metric to evaluate the two understanding tasks.

5. RESULTS AND ANALYSIS

This section is going to specify and analyze the experimental results. Based on previous work done by [6], we conduct continual domain-adaptive pre-training focusing on two relevant domains, moral culture and commonsense knowledge. [6] has post-trained RoBERTa [38] and T5 [21] on the unlabeled data and names them RoBERTa-Post and T5-Post in the original paper. Such post-training is the task-adaptive pre-training that we call in our paper, thus we rename them RoBERTa-T and T5-T in Table 5 for better distinguishment with our methods. The **T** in their names means **T**ask-adaptive pre-training, **TD** means both **T**ask- and **D**omain-adaptive pre-training, but the domain is moral culture. **TD+** means further pretraining about the domain of commonsense upon **TD**. **Human** means human performance on the two tasks, which has been tested by [6]. **#Para** is the approximate number of model parameters. For each task, the best performance is highlighted in bold and the second best is underlined, except for human performance.

Table 5. Accuracy(%) for CPT and PREF with different pre-training strategies.

Models	CPT	PREF	#Para
BERT [8]	59.62	82.97	110M
RoBERTa [38]	62.71	**89.54**	110M
RoBERTa-T [6]	64.61	<u>87.59</u>	110M
T5 [21]	69.60	82.00	220M
T5-T [6]	70.07	81.75	220M
T5-TD	<u>70.42</u>	82.68	220M
T5-TD+	**71.86**	82.41	220M
Human [6]	*95.00*	*98.00*	N/A

By analyzing the accuracy results in Table 5, we summarize our findings on two moral understanding tasks as follows: (1) T5 performs better than BERT and RoBERTa on CPT but worse on PREF, that tells that the encoder-only architecture might be good at aligning preferences. (2) We find that continual pre-training does not always imrpove the performance on

target tasks after comparing RoBERTa-T with RoBERTa and T5-TD+ with T5-TD on PREF, which advises that a better way is required to make use of these data especially when handling tasks similar with PREF. (3) We observe that different pre-training corpus brings different degrees of effects, which might depend on target tasks. T5-TD makes smaller progress than T5-TD+ on CPT, but the reverse happens on PREF, which indicating that the corpus of commonsense is more needed by CPT to enhance the ability of commonsense reasoning while PREF requires more moral data to capture value preferences. (4) Although a big gap exists between our models and human performance, continual pre-training has proved its effectiveness. Zero-shot or few-shot learning has been an important trend, which is supported by PLMs with strong generalization capability.

6. CONCLUSIONS

In this paper, we suggest to leverage a three-stage paradigm ("pre-training + continual pre-training + fine-tuning") instead of the traditional two-stage paradigm ("pre-training + fine-tuning"). The effects of the intermediate stage is tested on two downstream tasks of moral understanding. Specifically, the continual pre-training is categorized in two types, task-adaptive and domain-adaptive, with the aim of enriching the language model with task- and domain-awareness knowledge. Task-adaptive pre-training refers to further pre-training on unlabeled training corpus for target tasks before fine-tuning on labeled corpus. As for domain-adaptive pre-training, we utilize corpora from two different domains including moral culture and commonsense knowledge. To be specific, the corpus about moral culture is composed of Vernacular Chinese of Confucius theory. Furthermore, we linearize the triples of a Chinese commonsense knowledge graph into readable natural language sentences for incremental domain-adaptive pre-training. Experimental results reveals the effectiveness of our method, and requires paying attention to specific task property and the relevance between the domains and the target task. Continual pre-training performs better when the language model is more adaptable to the downstream tasks or when the content of the continual pre-training corpus is more supportive for them. Larger-scale pre-training over multitasks and multi-domains is of high computational cost but still necessary, especially in low-resource settings. For future work, we will figure out a better way to make the best of the corpora of continual pre-training, such as novel pre-training strategies and preferable data preparation.

ACKNOWLEDGEMENTS

This research was funded by the Research Development Fund at Xi'an Jiaotong Liverpool University, contract number KSF-A-17.

REFERENCES

[1] L. Jiang, C. Bhagavatula, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, R. A. Rini, and Y. Choi, "Can machines learn morality? the delphi experiment," 2022.

[2] Z. Li, X. Ding, and T. Liu, "Story ending prediction by transferable bert," in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[3] L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, "Plan-and-write: Towards better automatic storytelling," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7378–7385, Jul. 2019.

[4] F. Brahman and S. Chaturvedi, "Modeling protagonist emotions for emotion-aware storytelling," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Nov. 2020.

[5] X. Kong, J. Huang, Z. Tung, J. Guan, and M. Huang, "Stylized story generation with style-guided planning," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Aug. 2021.

[6] J. Guan, Z. Liu, and M. Huang, "A corpus for understanding and generating moral stories," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jul. 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Jun. 2019.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," ArXiv, vol. abs/1907.11692, 2019.

[10] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[13] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," in Proceedings of the 36th International Conference on Machine Learning, 2019.

[14] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general purpose language understanding systems," in Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., 2019.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," ACM Computing Surveys, vol. 55, pp. 1 – 35, 2021.

[16] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in SIGMOD Conference, 2008.

[17] D. Vrande˘ci´c and M. Kr¨otzsch, "Wikidata: A free collaborative knowledge base," Communications of the ACM, 2014.

[18] D. Li, Y. Li, J. Zhang, K. Li, C. Wei, J. Cui, and B. Wang, "C3KG: A Chinese commonsense conversation knowledge graph," in Findings of the Association for Computational Linguistics: ACL 2022, May 2022.

[19] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," in AAAI Conference on Artificial Intelligence, 2019.

[20] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in AAAI Conference on Artificial Intelligence, 2019.

[21] J. Guan, Z. Feng, Y. Chen, R. He, X. Mao, C. Fan, and M. Huang, "LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation," Transactions of the Association for Computational Linguistics, vol. 10, 2022.

[22] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, "A knowledge-enhanced pretraining model for commonsense story generation," Transactions of the Association for Computational Linguistics, vol. 8, 2020.

[23] F. Luo, D. Dai, P. Yang, T. Liu, B. Chang, Z. Sui, and X. Sun, "Learning to control the fine-grained sentiment for story ending generation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Jul. 2019.

[24] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 09 2019.

[25] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in The Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, 2020.

[26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020.

[27] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in Chinese Computational Linguistics, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Springer International Publishing, 2019.

[28] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in Proceedings of the 20th Chinese National Conference on Computational Linguistics, Aug. 2021.

[29] B. He, D. Zhou, J. Xiao, X. Jiang, Q. Liu, N. J. Yuan, and T. Xu, "BERT-MK: Integrating graph contextualized knowledge into pre-trained language models," in Findings of the Association for Computational Linguistics: EMNLP 2020, Nov. 2020.

[30] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "KEPLER: A unified model for knowledge embedding and pre-trained language representation," Transactions of the Association for Computational Linguistics, vol. 9, 2021.

[31] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Jul. 2019.

[32] Y. Liu, Y. Wan, L. He, H. Peng, and P. S. Yu, "Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning," ArXiv, vol. abs/2009.12677, 2020.

[33] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," Transactions of the Association for Computational Linguistics, vol. 8, 2020.

[34] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2018.

[35] Y. B. Ian Goodfellow and A. Courville, "Deep learning," MIT Press, 2016.

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Oct. 2020.

[37] P. Hosseini, D. A. Broniatowski, and M. Diab, "Knowledge-augmented language models for cause-effect relation classification," in Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022), May 2022.

[38] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," in Findings of the Association for Computational Linguistics: EMNLP 2020, Nov. 2020.

[39] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Nov. 2018.

AUTHOR INDEX